ARCHITECTS GUIDE TO

Fall 2016 Edition

IMPLEMENTING A DIGITAL TRANSFORMATION

Featuring the Big Data Maturity Model for IT

George Demarest
with Jim Scott

Architect's Guide to Implementing a Digital Transformation

George Demarest - author
Jim Scott - contributor

Table of Contents

Introduction

This document is meant to provide enterprise architects, IT architects, and other IT strategists some guidance as to how organizations can progress through the various stages of becoming a data-driven business. This guide describes four phases of the journey toward a digital transformation

The four phases:

Phase 1. Experimentation

- Understand capabilities of the big data ecosystem
- Develop basic skills in big data management and new application architectures
- Create a pilot use case
- Establish and maintain a working cluster

Phase 2. Implementation

- Develop the first production use cases
- Committ dedicated resources to big data development and operations
- Garner executive sponsorship
- Develop a plan for a broader digital transformation

Phase 3. Expansion

- Expand to multiple use cases across the company
- Plan participation by multiple lines of business
- Establish a dedicated command and control structure for digital transformation
- Establish IT SLAs, ROI metrics, and growth plan for data-driven operations

Phase 4. Optimization

- Optimize and integrate apps on converged data platform
- Establish digital business practices as the new normal supported by all key executive sponsors
- Provide detailed business SLAs, revenue targets, and other financial targets
- Normalize data lifecycle/governance, data monetization, microservice development

Where Did These Phases Come From?

These phases represent discernable patterns that we have observed from hundreds of engagements with MapR customers. While the phases are presented sequentially, individual experiences can vary significantly depending on the commitment of your organization, whether or not you have a motivated executive sponsor, external pressure from competitors in your industry, and other factors like budget, politics, and culture. MapR customer examples (both named and anonymized) will be used to illustrate key points of the big data journey.

The rapid growth of big data development is the result of major shifts in the broader information technology space. The combined impact of growing data volumes, accelerating adoption of open source technologies, and the trend in shortening the application development cycle have all contributed to the big data phenomenon. The rate of growth in the volume, variety, and velocity of data is accelerating. In the fixed Internet of the 90s, there were 1 Billion Internet connections. With the mobile Internet of the 2000s, it grew to 6

billion. And by 2020, according to Cisco there will be a projected 50 billion connections with the Internet of Things.

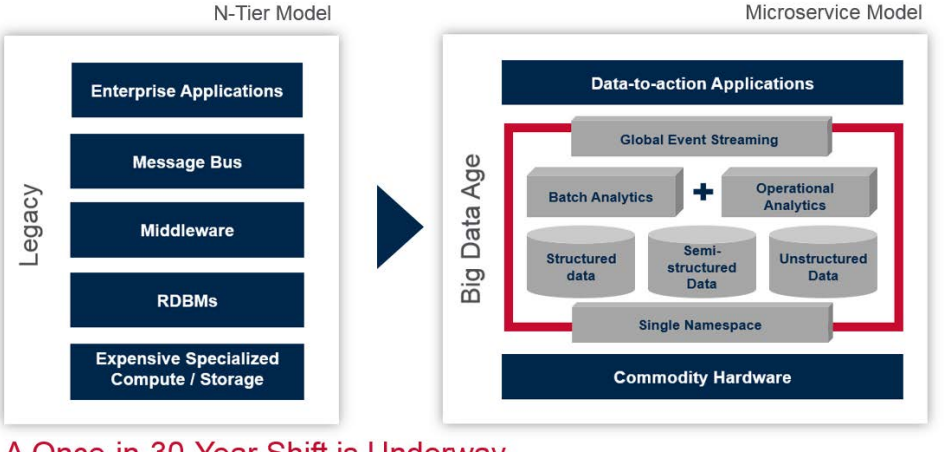
Digital Transformation, Open Source, and Cloud Economics

Legacy systems were never designed to handle this scale of data, yet a true digital transformation demands that you find a way. What many IT executives are discovering is that legacy practices, and more importantly, legacy economics, are being challenged by new digital platforms that exploit open source tools, and cost effective distributed computing. Couple that with the ability to develop these new applications on premises or in the cloud on virtualized infrastructure brings with it new financial models for IT that could be termed cloud economics.

More and more application architects and developers are asking questions like:

- Does my application really need a RDBMS?
- Is there a free or open source alternative to the commercial software I am using?
- Can I run a mission-critical application without any commercial software?

In our own customer base, there are dozens of examples of re-platforming of applications or analytics from legacy platform such as mainframes, data warehouses/RDBMS, and premium storage arrays. The Hadoop/Spark ecosystem first grew in popularity because it provided an economical way to store massive amounts of data and do bulk processing on these large data sets.



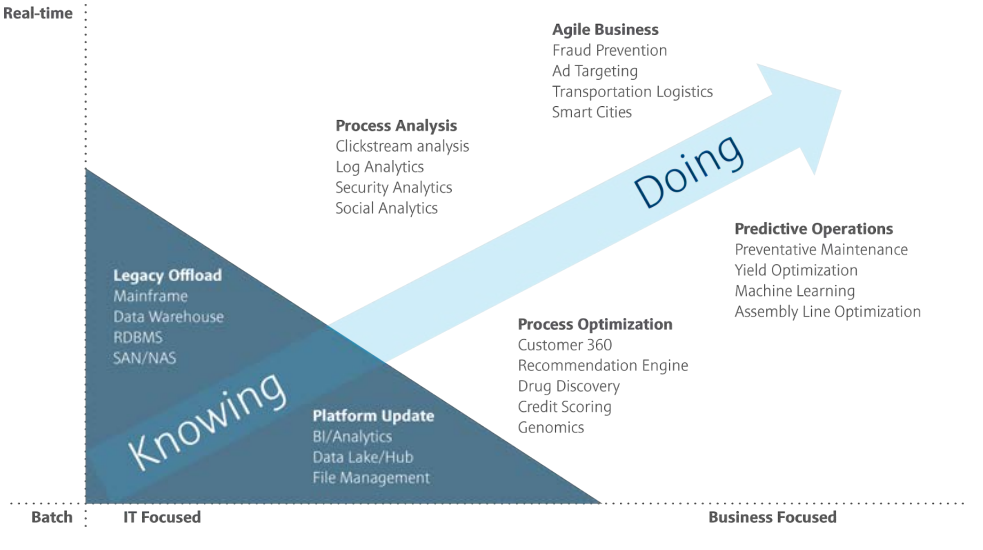
A Once-in-30-Year Shift is Underway

This proliferation of new data, new tools, and new thinking offers organizations tremendous opportunities to reach and serve customers through new application architecture. The growing utility and influence of machine learning, artificial intelligence, advanced analytics, data engineering, statistical analysis, and data science vastly expand the lexicon of "business intelligence". The impact of such a sea change in approach to data-driven business processes in the digital business suggests the ability to reimagine the role of IT in the business.

The Progression of Big Data Use Cases

While every organization has a unique entry point into the world of big data, we have been able to observe some fairly consistent patterns of use case development from our customer engagements. During the early stages of big data adoption, IT departments sensibly focus on easy wins that focus on cost savings, gaining valuable practical experience, and laying the groundwork for more ambitious and sophisticated projects.

The graphic below roughly outlines how use case development progresses in most customer environments. There are, of course, exceptions. Typically, those exceptions stem from the customer having a critical use case or competitive situation that accelerates adoption of the technology. Use cases such as security analytics, fraud detection, marketing use cases, or Internet of Things (IoT) projects can fall into this category.



While use cases are a key measure of the maturity and sophistication of a digital transformation, this document presents a broader set of indicators that are meant to inform enterprise architects, IT leadership, and application architects. While it is difficult to generalize the economic benefits of big data or the MapR Converged Data Platform, we have hun-

dreds of examples of customers achieving significant — sometimes radical — costs savings, new revenue streams, and high returns on investment. See the **IDC report**.

Finally, the fast pace of change in the big data ecosystem, the great strides being made in data engineering and data science, and the growing number of "data-driven" business and IT leaders mean that the phases described below will also evolve over time. This current version is based on conditions on the ground today and is based completely on real data and experience from MapR customers to date.

Phase 1: Experimentation

Understanding the capabilities of the big data ecosystem

In the first phase of the big data journey, companies are exploring how the Hadoop ecosystem works and how it can fit in with their existing enterprise data architecture. For example, scaling the business has become a cost or performance hurdle which requires a new approach. Their legacy systems can't handle the volume or variety of new data being generated. They are spending a lot on additional storage and not putting this extra data to work, and there are multiple data silos that don't work together.

Motivating Factors

Cost takeout

The combination of low cost commodity servers with cheap direct attached storage as well as open source software, provides an enticing reason to explore big data projects. Interestingly enough, we find that there is a general increase in CPU power, memory, and storage footprints for server nodes in MapR installations. This seems to stem from the high levels of performance of the MapR File System and other converged components. However, as you will note below, some of the most significant cost savings come from the retiring of expensive software license and support costs, and premium hardware used in legacy BI/DW and mainframe environments.

IT preparedness

Even companies at early stages of digital transformation will have a CIO that looks at least three years out. IT may or may not drive/control every aspect of digital business processes, but they do control their own readiness to implement data-driven technologies when they are called upon to make digital transformation a reality.

Initiate IT-wide data strategy

It is universally accepted that much of the total collected data in any given organization is materially valuable intellectual property, and is growing in value due to new ways to an-

alyze and act on the data being collected. Therefore, many organizations are moving to rationalize and then monetize internal data sets and to explore augmenting that data with external data sources such as weather data, credit and finance data, trading data, and so on.

Key Activities

Identify a team

Any successful big data project depends on the focus of those running it. At the earliest stages, all of the critical resources (developers, hardware, sysadmins) will likely be on borrowed time. Time-shared or off-the-books hardware, half-time developers, and administrators are often how these projects are initiated. Moving forward, the creation of a more formal big data Center of Excellence (COE) is a good early goal. That COE would include distinguished staffers from IT Ops (for cluster admin, data management), members of the BI/DW/Analytics team, and perhaps the most forward-looking developers.

Skills and tooling update

Plan for the core COE team to invest in some training to prepare them to do real work on the applications and the underlying platform. MapR does provide some support here with an extensive, free **online curriculum** and professional certifications for Hadoop, Spark, the MapR Converged Platform, cluster administration, and other related topics. It can provide a significant head start for your teams and is rightly considered a significant contribution to the Hadoop/Spark community.

Identify systems or applications to offload/migrate

As referenced above, offloading legacy systems is an attractive target for early stage big data projects. A sensible first step to take in this area is to begin to prioritize key systems that need modernization or whose cost models don't scale. Doing this type of investigation can often yield valuable information needed to get your digital transformation off the ground. MapR offers a tool to calculate **total cost of ownership** for a MapR solution that may be useful in this regard.

Plan for growth

Even when experimenting, it's important to make the right decisions up front to prepare you for growth in the future. For example, you may end up wanting to use MongoDB or Kafka, so be mindful of new technologies and how you will integrate them in the future.

Use Cases

Use cases in Phase I can be characterized as basic and limited in scope. That is not to say they are not useful or simple. But in any learning process, you take initial steps that are clear and verifiable and then add sophistication and more ambitious goals as you progress.

Platform bake-offs

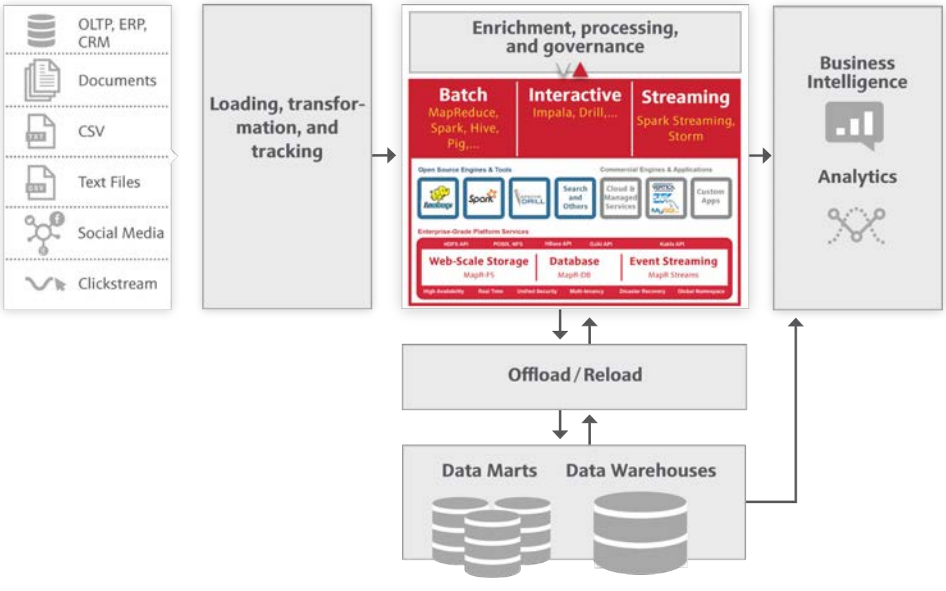
The most basic and obvious activity in coming up to speed on the complex set of technologies that make up the Hadoop/Spark/Analytics ecosystem is to compare individual vendors and open source distributions.
Clear targets for evaluation include:

- Hadoop distributions - MapR Converged Data Platform versus Cloudera, Hortonworks
- Spark distributions - MapR Platform Including Spark, versus Data Bricks (cloud), Hortonworks
- Streaming - MapR Streams versus Apache Kafka, TIBCO, RabbitMQ
- NoSQL Database - MapR-DB, versus MongoDB, Cassandra, HBase
- Analytics - Spark core, machine learning, in-memory technologies, updated BI platforms like SAS, Qlik, Microstrategy
- SQL on Hadoop - Apache Drill, Impala, Spark SQL, Hive

Legacy offload

Another common starting point on the road to big data is the process of offloading data or processing from legacy systems. This is a sensible move in that a successful offload can mean reducing license costs, reducing support/maintenance costs, as well as cost avoidance (for instance, if you can halt the expansion of a DW investment). Most likely targets are:

- Data warehouses - by either offloading ETL or offloading cold data
- Mainframes - costs savings can be significant on both storage and processing
- Storage systems and file servers - the inherent triplicate replication of the MapR Platform provides ample protection for file data



Some estimate that implementing a Data Warehouse with one of the leading vendors starts at $1 million. Many are finding that the MapR Platform can accomplish many DW tasks at a fraction of that cost. The “return on divestment” can end up funding the entire big data program.

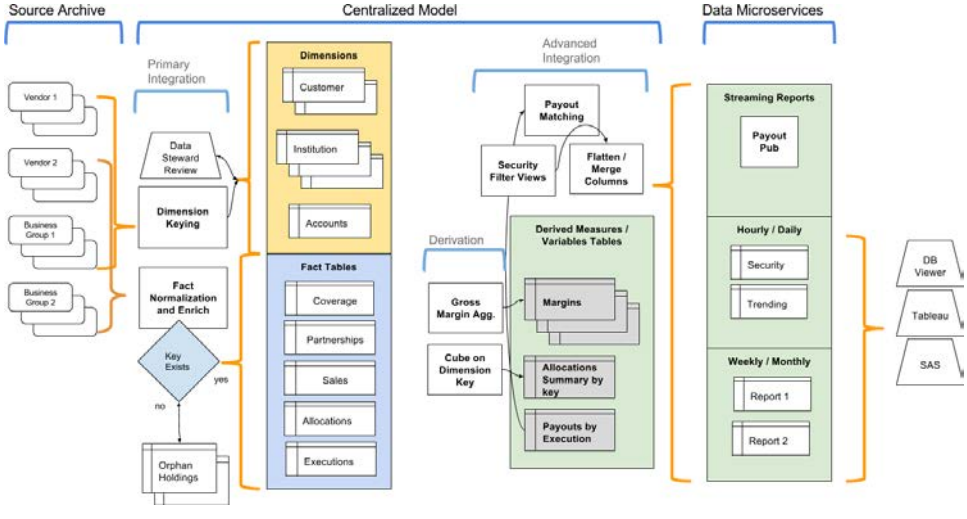
Data Warehouse Offload

Data Warehouse offloads (DWOs) are one of the most common use cases among MapR customers. The MapR Professional Services team has helped dozens of customers through this process. Here is a high level overview of their methodology that they describe in three steps:

1. **Creating the Source Archive:** As a practical matter, the process for offloading a legacy data warehouse to Hadoop, or better yet the MapR Platform, is one of extracting data into the Hadoop environment to create a Source Archive, basically raw data without transformations, integrations, or cleansing. In such a format, your ability to query data is limited to schema-on-read technologies and performance may not be great.
2. **Build a Centralized Data Model:** The next step is to begin to create data model for all data sources. This data model is not yet optimized for speed, but is created using a methodical approach to ensure better resource management, better data quality, better change control procedures, and a better understanding of the data lineage.

3. **Establish Data Microservices:** The final step is to do more and more derivations, aggregations, and amalgamations in order to address the direct needs of users/customers by delivering *data microservices*.

DWO Offload Workflow:



Data Warehouse Offload Use Case

One of the most common first projects with Hadoop and the MapR Converged Data Platform is to extend the life and lower the costs of existing data warehouses.

Management Science Associates (MSA) Case Study

Driving Organization: IT Operations and Data Warehouse team

The Business

Management Science Associates, Inc. (MSA) provides point of sale analytics to help consumer packaged goods companies improve marketing programs and in-store presence. MSA collects and processes data from 7,000 distributor-warehouse files, reporting its shipments to over one million unique retail outlets. MSA analyzes this data to help their customers detect demand for a product so distributors and manufacturers can automatically adjust orders to meet demand.

The Challenge

As the volume of data was increasing, MSA was facing rising costs when trying to scale their legacy Oracle database and storage infrastructure. And their customers wanted a faster response time to queries.

Solution

MSA is building a next-generation data warehouse environment designed to provide customers with near-real-time predictive analysis.

Benefits

- **Multi-tenancy** lets MSA keep distinct data sets and customers isolated from each other—all in a single consolidated deployment.
- **Integrated NoSQL** provides customers with access to structured and unstructured data.
- **Disaster recovery** and high availability is critical to their business. They run two data centers, and having the ability to fail over between them is a critical criterion.
- **Performance** increased with the MapR File System so they are able to use less hardware and get better performance.

Further Information on Legacy Offload:

- *MapR Solution: **Data Warehouse Optimization***
- *Case Study: **Management Science Associates Builds Next-Generation Data Warehouse and Near-Real-time Analytics on MapR***
- *Case Study: **Goodgame Studios Takes Gaming to the Next Level with MapR***
- *Case Study: **Valence Health® Dramatically Improves Data Ingestion Performance and Scalability with MapR***
- *Case Study: **Mainframe Offload: Experian Increases Insights and Speed to Market with MapR***

IT focused

IT organizations can also focus their attention on their own systems and data thereby allowing them to build skills and confidence in using new technologies without outside scrutiny or pressure. It also provides a means for IT operations teams to come up to speed on the technologies independent of a bespoken program brought about by a line of business or software development teams.
Some examples include:

- User home directories
- Machine log analytics (trial)
- File management
- Cold data offload/archive

Log Analytics, Machine Learning Use Case

Arvest Bank

Driving Organization: Network Security

The Business

Arvest Bank provides a wide range of financial services including loans, deposits, treasury management, asset and wealth management, life insurance, credit cards, mortgage loans, and mortgage servicing. The bank is owned and controlled by the Walton family of Walmart.

The Challenge

The bank was vulnerable to cyber attacks and might not meet SEC and FINRA requirements with its current Security Information and Event Management (SIEM) at capacity. They had over 2500 sources feeding the SIEM, and required every Windows and Linux server to send their security logs to it as well. Their existing SIEM was hitting capacity. They were also struggling with how to manage semi and unstructured data. They need to be more "real-time" in responding to customer data breaches.

Solution

They completely moved off SIEM with one central, real-time MapR Platform for protecting customer data. They chose MapR for high availability and high performance

Benefits

- 46X cost savings over alternate solution. MapR was $60K/year vs. $2.88 million/year

- SEC and FINRA requirements met
- Removes silos with all sources and complexity
- Success will be measured on customer satisfaction and lowered attrition

Checklist to Progress to Phase II: Implementation

Here are areas you should be thinking about as you move into phase 2. You should be playing around with all of the technologies.

Organization

- Do you have a plan to export your data?
- Do you have multiple people who can learn the technology?
- How do you plan to train your team?
- What's your comfort level in each area?

Operations

- How do you get data into your system?
- Do you understand backups and recoveries?
- Do you have a disaster recovery plan?

Development and BI

- Do your developers have familiarity with Spark and MapReduce?
- Have you considered a SQL-on-Hadoop strategy?

Summary

Phase I: Experimentation	
Description	Understand capabilities of big data platform
Motivation	Cost takeout IT preparedness Initiate big data strategy Develop staff skills Evaluate alternative technology

Executive Sponsor	**One of:** CIO CTO VP of Enterprise Applications VP of DW/Analytics CEO CMO CRO
Staffing	**Part time:** Cluster administrator Developer BI analyst
Participating Organizations/ Groups	Central IT (infr) Application Development BI/Analytics LOB IT
Hardware Investment	Public Cloud Servers on Premises Borrowed/shared Near end of life (commonly done on off-the-books hardware)
Number of Nodes	2-5
Key Capabilities	Easy installation Easy ingest Application templates, scripts
Key Technologies	Easy ingest (NFS) File System NoSQL Database
Key Skills Required	**IT operations:** Basic software/cluster install and data management Manual data ingest Basic Hadoop administration Linux system administration **Software development/BI:** **Batch:** MapReduce **Interactive:** SQL on Hadoop (Drill, Impala, Hive) **DWI/BI/Analytics team:** Basic NoSQL Log analytics Basic statistical skills (R)
# of Production Use Cases	0-1
Common Use Cases	**Platform bakeoffs:** Hadoop Event streaming systems NoSQL databases Analytics SQL on Hadoop

	Legacy offload: Data warehouse Mainframe Storage (NAS/SAN) **IT focused:** User home directories Machine log analyics File management Cold data offload/archive
Number of Data Sources	1-5
Data Sources	Legacy servers and strorage (Files) Date Warehouses/RDBMS File Servers, SAN/NAS Sample Data

Phase 2: Implementation

Developing the first production use cases

In the Implementation phase, companies are ready to get started in creating production use cases. They should have a comfort level and confidence that they can successfully build a solution on this platform. They have enough proof (successful pilot, ROI metrics, working programs/dashboards) to show IT management that the program is viable. They want to choose an important use case for the business, build it, test it, and prepare it for production use. They will be working to fully operationalize the platform, develop a mechanism to support it every day, and then move it into production.

Motivating Factors

- **Creating a big data discipline** - moving beyond the proof of concept/value stage, big data proponents must take steps to establish key platform components (like Spark, Hadoop, Solr, ElasticSearch, etc.) as permanent parts of their technology portfolio.
- **Delivering measurable benefits** - key to continued success of a big data program is the successful execution of one or more use cases that prove the viability of big data and advanced analytics. If a big data program is to become a permanent dimension of your IT strategy, then objective, provable metrics as to service levels, budget, and efficacy are required.
- **Staff satisfaction and retention** – many in IT operations or development are strongly motivated to maintain up-to-date skills in leading edge tools and technologies to keep their careers on track. Open source technologies around big data, analytics, and adjacent technologies — like micro-service development, containers, streaming analytics, and IoT — are high on the list of what enterprise developers and those in IT ops want to be working on.

Key Activities

Execute lighthouse win. The big data program team must drive a small number of use cases into production that delivers a demonstrable benefit to the business. These early use cases will set the tone for the entire program going forward.

Use case expansion: ideation and roadmap. Once the big data program team has demonstrated significant success with their initial use cases, there should be no problem expanding the list of potential use cases that build on early efforts. Due thought should be put into both cost savings and revenue generation when considering future use cases.

Develop a longer term business plan that includes:

1. Progressive cost takeout strategy for legacy systems (RDBMS, middleware, data warehouses, mainframe, premium storage)
2. Top-line revenue-focused use case strategy, with stretch goals of business disruption
3. Skills roadmap for IT Ops, developers and your BI/DW team
4. Support strategy for open source platforms and tools

Use Cases

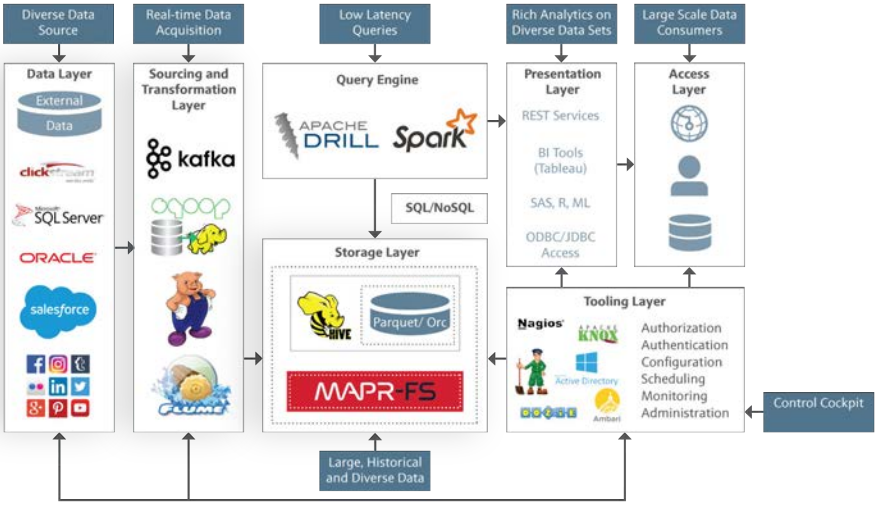
Regardless of which use case looking at, starting to create your own data lake/hub. Now is the time to plan for multiple use cases coming along to run on the same data.

Data lake

The **Gartner definition of a data lake** is

> ...a collection of storage instances of various data assets additional to the originating data sources. These assets are stored in a near-exact, or even exact, copy of the source format.

For a majority of MapR customers, the data lake is often their first use case and its construction is a key part of the process of establishing big data technologies and practices in an IT organization. By its nature, it is meant to span organizational or technological silos, which introduces some interesting new possibilities. The data lake is often an essential first step to a customer/citizen/patient 360 use case.



The primary function of a data lake is to assess and rationalize data sources, thereby dismantling "data silos" that keep analysts from getting a complete picture of their operations and their customers.

Data Lake/Data Hub Case Study

Razorsight/Synchronoss Case Study

Driving Organization: IT Operations

The Business

Razorsight's cloud-based predictive analytics software delivers insights to help communications service providers (CSPs) and media companies proactively improve customer experiences, reduce costs, and increase margins.

The Challenge

Today's telecom data has higher volumes, frequency, and more complex structures. There are new types of devices generating data for the Internet of Things, mobile phones using broadband for apps, and VoIP. Razorsight had to evolve its technology stack to achieve scalability at a reasonable cost.

Solution

Razorsight used MapR to build a central data lake as a primary data store for both online and archive data. The data lake ingests customer data in all shapes and formats from multiple sources. Since the launch of this new stack in late 2014, the production cluster has received, processed, and analyzed more than 40 terabytes of data.

Benefits

Improved performance. With their previous architecture, Razorsight ran into multiple bottlenecks because data ingestion, processing, analytics, querying, and visualization were all competing with each other for processing power. With the MapR platform, they can completely separate these, resulting in a huge impact on performance and scalability.

Cost savings. The MapR architecture also results in significant cost savings. The total cost of storage and processing for a traditional enterprise EDW platform is about $15,000–20,000 per terabyte. With the Hadoop ecosystem, this has dropped to about $2,500–3,000 per terabyte for them.

Easy integration. Razorsight leverages the MapR NFS gateway to move data sets in and out of the cluster seamlessly, making it extremely easy and intuitive to integrate Hadoop into the overall data flow.

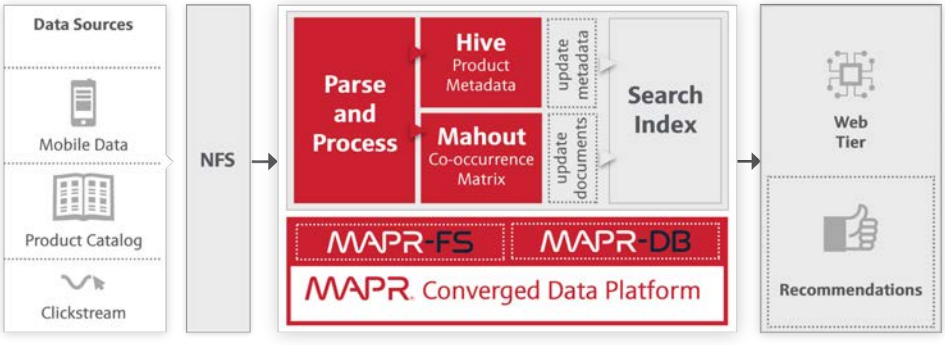
Further Information on Data Lakes:

- *MapR Solution:* ***MapR and Data Lakes***
- *eBook:* ***The Definitive Guide to Data Lakes by Radiant Advisors***
- *Case Study:* ***HP Leverages the Power of MapR in its Big Data Infrastructure***
- *Case Study:* ***How Cisco IT Built Big Data Platform Using MapR Distribution to Transform Data Management***
- *Case Study: MAG45* ***MapR Data Lake Solution Cuts New Customer On-boarding Time by 50%***

Marketing

Marketing functions have become some of the most common big data use cases for good reason. Any business that relies on communications, awareness, selling, and other customer interactions have various types of market problems to solve. Specific marketing use cases include customer 360, recommendation engines, churn analysis, ad and content targeting, next likely purchase, sentiment analysis, and many more. With the CMO much

more interested in data and technology than ever before, marketing use cases are often funded first because they can present clear return on investment.



Further Information:

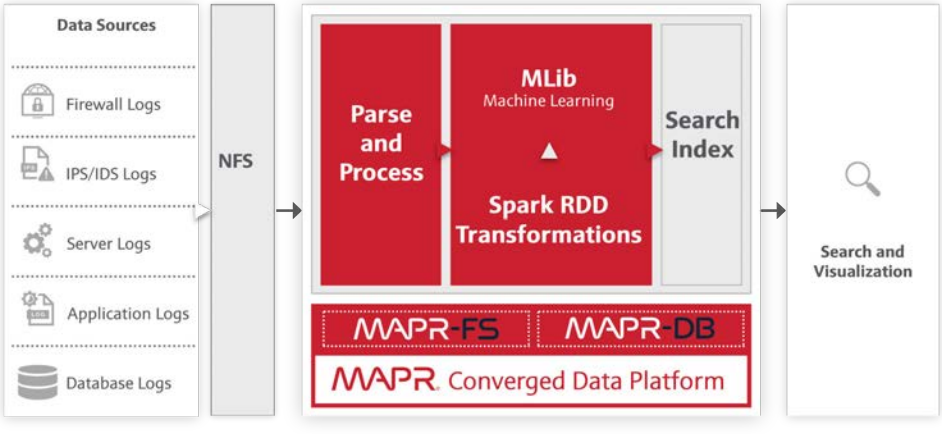
- *Case Study:* ***Harte Hanks Uses MapR Platform to Target Customers Faster and More Accurately***
- *Case Study:* ***Rubicon Project : Leading Digital Advertising Auctions Platform***
- *Case Study:* ***IRI Builds Cost-Effective, Scalable Platform for Market Intelligence and Analytics***

Security and Fraud Detection

Like marketing use cases, there is a broad appeal for security-related use cases for both security tech vendors and the broader market with security, privacy, fraud, and intrusion detection concerns.

- **Security Information and Event Management (SIEM)**: Analyze and correlate large amounts of real-time data from network and security devices to manage internal and external security threats, improve incident response time and compliance reporting.
- **Application Log Monitoring**: Improve analysis of application log data to better manage system resource utilization, security vulnerabilities, and diagnose or preempt production application problems.

- **Network Intrusion Detection**: Monitor and analyze network traffic to detect, identify, and report on suspicious activity or intrusions.
- >**Fraud Detection**: Use pattern/anomaly recognition on larger volumes and greater variety of data to detect and prevent fraudulent activities by internal or external parties.
- **Risk Modeling**: Improve risk assessment and associated scoring by building sophisticated machine learning models on Hadoop that can take into account hundreds or even thousands of indicators.



Security Analytics Use Case

Solutionary Case Study

Driving Organization: IT or CISO

The Business

Solutionary delivers managed security services and professional consulting services, or threat intelligence as a service to its clients. The company uses proprietary security analytics technology to reduce risk, increase data security, and support compliance initiatives for its clients.

The Challenge

They needed to improve scalability as the number of clients and data volume grew but it was cost-prohibitive with their existing Oracle database solution. The old solution could not process unstructured data at scale and there were also major performance issues.

Solution

They replaced their RDBMS with MapR to achieve scalability while still meeting reliability requirements. Their new solution combines machine learning algorithms, complex event processing, and predictive analytics to detect real-time security threats.

Benefits

- Reduced time needed to investigate security events for relevance and impact
- Improved data availability, enabling new services and security analytics
- Enhanced agility with ability to deploy on-demand capacity
- Superior performance enables far more complex processes. Solutionary can now globally detect and analyze threats across all clients within milliseconds.

Further Information:

- ***Real-time Security Log Analytics with Spark on Hadoop***
- ***Video: Security Analytics and Big Data: What You Need to Know***
- ***Solutionary: Leading Pure-Play Managed Security Service Provider***

SaaS Architecture and Enterprise Application Replatforming

Re-engineering an existing enterprise application is typically a non-trivial task, but it is not uncommon for SaaS providers who establish a baseline of functionality and a measure of success. Sometimes a symptom of that success is that the original architecture has scalability issues either due to high data volumes or a monolithic processing design. Another reason to replatform might be for economic/budgetary concerns.

Relational Database Take-out (RDBMS -> NoSQL)

Driving Organization: Database and Data Warehouse teams

The Business

This financial data and software company provides financial information and analytic software to investment professionals.

The Challenge

The company's database includes approximately 100 mutual fund holdings. Anytime there is a change in a holding, position, or change to a stock, it is updated in the database, and then they deliver services back to investment professionals.

Solution

They wanted to transform their 20-year-old legacy VMS platform to MapR-DB and they created a new data platform.

Benefits

- Scalability and high performance
- Multi datacenter replication
- One data platform moving forward

Further Information:

- *Case Study:* ***Valence Health Builds New Data Architecture on MapR to Keep up with Growth***
- *Case Study:* ***sovrn Uses MapR as Foundational Data Platform for Online Advertising Exchange***

Checklist to Progress to Phase III: Expansion

Here are areas you should be thinking about as you progress to phase 3.

Organization

- Have you identified the new lines of business that can benefit from new big data use cases?
- Have you identified and gained the resources (staff, hardware, software) you will need to expand your program?

- Have you defined KPIs and other metrics to demonstrate the ROI of current and planned production use cases?
- Have you developed a detailed business plan to justify the expansion of your program?

Operations

- Do you have a documented and measured means of deploying, supporting, and managing the application lifecycle for big data platforms and use cases?
- Do you have a security, privacy, governance, regulatory standards, and processes designed for the new platform?
- Do you have a clear method of measuring service levels that are meaningful given your current standards and practices?

Development and BI

- Do your developers and BI staff have the skills and knowledge to exploit the new development paradigm for big data use cases?
- Have you designated a portfolio of primary tools and technologies that will make up the bulk of your digital platform?
- Have you identified resources (internal or external) to provide guidance and expertise in data science, data engineering, application architecture, and use case discovery?

Summary

Phase II: Implementation	
Description	Develop first use cases and put into production
Motivation	Creating a big data discipline Delivering measurable benefits Staff satisfaction and retention
Executive Sponsor	**One or more of:** CIO CTO VP of Development/Enterprise Applications VP of DW/Analytics LOB Executive CMO CISO CRO

Staffing	**Dedicated:** Cluster administrator Developers (1-3) BI analysts **Part Time:** Dev ops Data engineer **Optional:** Staff augmentation (consultants/contractors) Data engineer Data scientist
Participating Organizations/Groups	Central IT (infr) Application Development BI/Analytics Lines of Business (1-2) Marketing Sales Security
Hardware Investment	Dedicated cluster Public cloud
Number of Nodes	3-6
Key Capabilities	High-throughput, frictionless ingest Enterprise-Grade Persistence/Storage Basic Security Authorization Access Controls High Availability Replication Failover Mirroring
Key Technologies	NFS MapR-FS MapR-DB Apache Spark (core) Apache Drill
Key Skills Required	**IT operations:** Data management Data ingest Hadoop administration Offload/extend life of legacy Storage offload Ad-hoc data engineering **Software development:** **Batch:** Hadoop ecosystem **Interactive:** Spark **Query:** SQL on Hadoop **Search:** Solr, ElasticSearch Microservice pilots **DWI/BI/Analytics team:**

	NoSQL Log analytics Schema-less ad-hoc query (Drill, et al) Expand statistical skills (R)
# of Production Use Cases	1-5
Common Use Cases *from previous phase	Legacy offload: Data warehouse Mainframe Storage (NAS/SAN) IT focused: User home directories Machine log analytics File management Cold data offload/archive Data lake/hub: Analytics Platform Data Platform Application replatforming: Date Warehouse retirement RDBMS application re-engineering Reengineer legacy apps **Marketing/Sales:** Recommendation Engine Customer 360 Next likely purchase Customer churn Ad/content/customer targeting **Security:** Security log analytics Fraud detection **Operations:** IoT/Industrial Internet Supply Chain optimization/analytics Logistics Predictive/preventative maintenance
Number of Data Sources	5-15
Data Sources	IT systems Date Warehouses/RDBMS File Servers, SAN/NAS Application/System Logs Clickstream ERP/CRM systems

Phase 3: Expansion

Expand to build multiple use cases across key lines of business

Success with early use cases demonstrates to others in the company what can be accomplished. Expansion of the use cases from earlier phases is a natural next step. Implementing them will be a little easier since the platform is in place, but if the use cases cross team boundaries or are considered interdisciplinary, then multi-tenancy, security, and data governance will become a concern.

In building new use cases and applications, attention now turns to the many questions that arise from scaling any project:

- Are you ready to handle the scalability of those applications within your organization?
- Are there multiple open source projects, cluster paradigms, and data formats in your solution?
- Do they talk to each other? Do you monitor them?
- How do you scale their communication?
- If you scale one tier are you required to scale the other tiers?

These type of questions lead to application architecture and deployment architecture discussions of topics such as microservices development, event streaming, data pipelines, governance models, and application/data lifecycle.

Motivating Factors

Establish and meet IT service level targets and ROI goals

IT organizations have by now normalized their IT processes for Hadoop, Spark, and other associated big data ecosystem components. Standard service level metrics should exist for security, uptime, performance, resource usage and management, governance, etc.

Goals and metrics should also be established cost savings (CAPEX/OPEX), operational efficiencies, plans, and ideation for new revenue generating opportunities.

Establish success for key lines of business

That is acknowledged by LOB leadership such as:

- **CMO** - sales and marketing use cases
- **CIO** - improved IT operations, executive dashboards, legacy offload
- **CISO** - risk, fraud, and security intelligence
- **COO, GM** - significant progress should be made on vertical specific use cases

Prepare for the post big data digital Business

Organizations should be in the process of establishing a permanent Data Engineering and Data Science capability. IT operations teams should have redundant coverage on cluster administration for Hadoop and Spark. BI teams should have developed skills in predictive analytics, machine learning, event processing, and streaming analytics.

Key Activities

Establish standard big data application and analytics platform

Includes:

- Budget, staffing, train, big data teams in Central IT, BI/DW, software development, and related SME leads in participating lines of business
- Consolidate any Hadoop cluster sprawl. Often Hadoop projects begin with grassroots development in a business unit/area and as the business wants to have IT support the system in an on-going basis, IT looks to consolidate vendors/clusters for better TCO and reusability

Rationalize production use cases

By this phase, you should have between 2-10 production use cases at the start of the cycle.

- All possible legacy systems and applications re-deployed, offloaded or optimized
- Make your data lake/hub a 24x7 business application which meets business SLAs
- Begin to offer big data shared services to LOB
- Develop opportunistic apps - often co-driven with LOB leaders
- Deploy 2 - 10 apps/use cases (some/all in production with SLAs)

Plan to operationalize applications, analytics, and insights

- Identify “operational” applications which can take advantage of these new insights
- Plan development activities to take advantage of analytics to change business processes and business outcomes

Considerations for Expansion

Don’t defeat your data lake

You should continue to keep track of a lot more valued information from many more sources, for longer periods of time. Maintaining such extensive amounts of historical data presents new opportunities for deeper business insights.

Avoid cluster sprawl

You should take care not to allow new tools and technologies to force you into establish often from startups or other cutting-edge vendors. However, many of these new solutions are operating in isolation from the rest of your IT portfolio.

Define advanced data management

To achieve the highest levels of use case sophistication, performance and analytical accuracy, IT organizations should take advantage of volume management — (information lifecycle management, data governance best practices.)

Operational concerns as you expand your production footprint

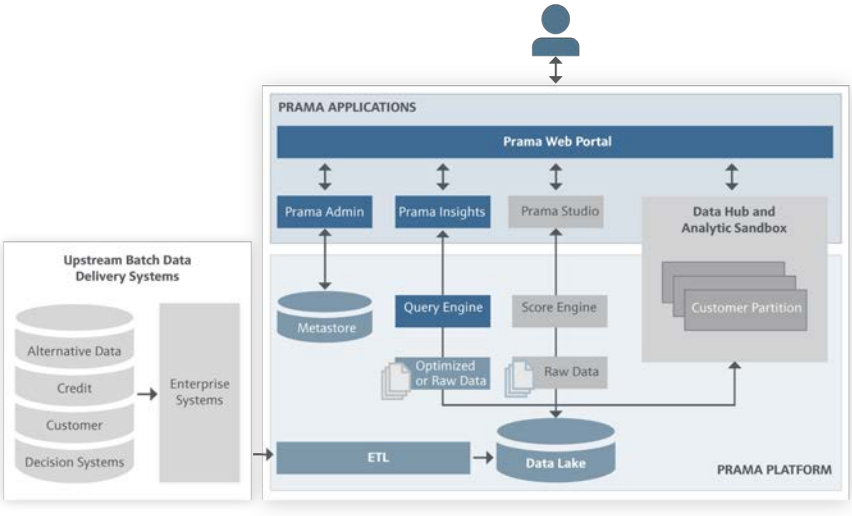
As you expand your big data program, you may be tempted to create multiple clusters to isolate lines of business, particular data sets, or applications and use cases. This is not recommended as it reinforces “legacy thinking” where organizations want to maintain ownership of individual hardware or other IT resources. But creating multiple separate clusters (e.g. Hadoop, Spark, NoSQL, Kafka, etc.) establishes a new set of data and operational sATLAS-CURSOR-HEREilos.

Use Cases

The expansion phase is named as such because of what logically follows initial production use case. The success of early use cases is often followed by a flurry of new use cases in related areas. Thus, a single marketing use case is followed by a marketing suite or practice. Security analytics becomes a SIEM suite. Thus, the focus becomes less on individual use cases and more on how use cases are combined to create a business platform.

Data/Analytics Platform, Analytics as a Service

While it may seem like there's a fine line between a data lake and a data or analytics platform, there is an important distinction. As stated above, the data lake is usually an early stage use case that focuses on rationalizing data silos and surfacing the complete wealth of data available to BI teams and business analysts. But when a data lake evolves into a platform that provides analytics as a service to internal users or external customers, then you begin to enter the realm of Analytics as a Service. An example below is TransUnion and their Prama analytics platform.



Further Information:

- *Case Study:* ***comScore Reliably Processes Over 1.7 Trillion Internet and Mobile Events Every Month on MapR***
- *Case Study:* ***TransUnion Launches New Self-Service Analytics Platform with MapR Technology***
- *Case Study:* ***Pico Selects MapR as Technology Foundation to Develop Innovative Services***

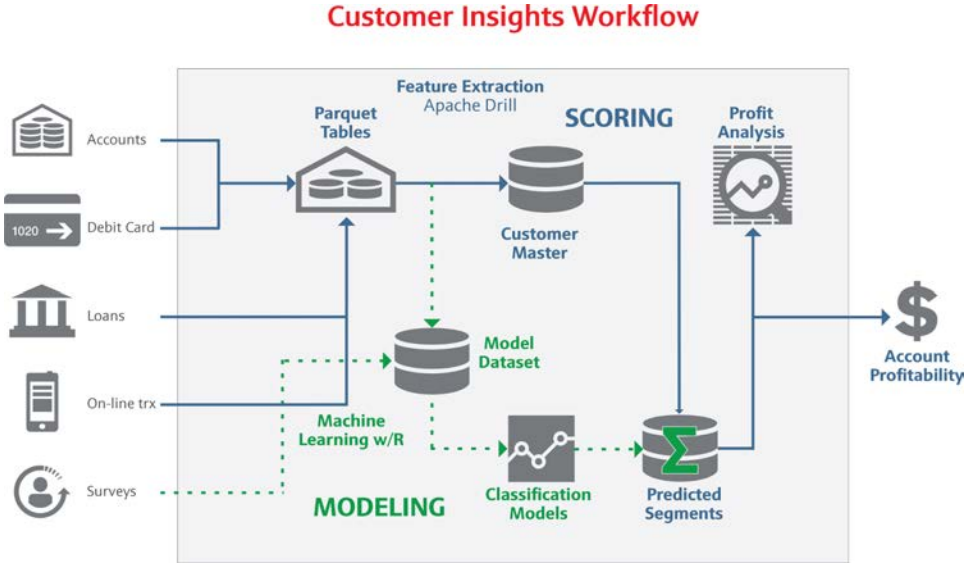
Marketing/Sales Suite

As detailed in the implementation phase, marketing use cases like recommendation engines and customer 360 are common candidates as first production use cases, especially in industries like ad-tech, ecommerce, and online retailing. As CMOs and their marketing analytics teams expand from their original use case, they quickly move to a set of standard marketing processes that become part of standard practices.

That suite would include, but not be limited to:

- *Customer 360*
- *Recommendation engine*
- *Next likely purchase*
- *Ad/content targeting*
- *Social/sentiment analysis*
- *Clickstream analytics*
- *Customer experience*

Below is an example implementation architecture for customer 360.



Customer 360 Case Study

Hewlett-Packard

Driving Organization: Big Data Services IT Team

The Business

HP needed to find the right technology to build an infrastructure for its internal big data development projects and to serve as the platform for new client offerings. HP wanted to increase customer satisfaction by providing a consistent, knowledgeable real-time customer experience.

The Challenge

- Customer information was siloed in different divisions
- Customer interactions were inconsistent and not satisfying
- There were missed opportunities for upselling and cross-selling

Solution

HP now integrates all customer data from across the company into a central data repository based on MapR technology. All divisions can access a customer dashboard that includes all interactions that each customer has with the company.

HP chose MapR for its superiority to other solutions in performance, high availability, disaster recovery, manageability, knowledge base, and future roadmap.

Benefits

HP leverages MapR as a low-cost, massive storage platform to integrate, consolidate, and analyze data from multiple sources. The HP data lake has enabled the development of new solutions and client offerings, which are helping to improve the overall HP customer experience across all touchpoints.

Security Suite

Like many other highly technical, highly specialized disciplines, the dev teams that create cyber security and fraud detection use cases quickly identify new related or expanded use cases that begin to form the foundation of a Security Information and Event Management (SIEM) system.

The Security Operations Centers (SOCs) often become the first "switched on" element of a digital transformation; especially since its failure could potentially bring the whole thing crashing down. Due to the nature of cyber security and cyber crime, the components of a

security suite of use cases must be updated constantly. Here are just some of the use cases that are in production at MapR customers:

- *Security log analytics*
- *ATM fraud*
- *Insurance claims fraud*
- *Anti-money laundering*
- *Advanced threat/intrusion detection*
- *Secure data vault*

Fraud Detection Case Study

UnitedHealth Group (UHG)

Driving Organization: Payment Integrity Group

The Business

This $11B diversified managed health care company serves more than 85 million people worldwide with health benefits and services. The Payment Integrity group is responsible for ensuring that claims are paid correctly and on time.

The Challenge

UHG wanted to reduce fraud and waste in payments and increase efficiency of the claims processing process. Their previous approach to managing more than one million claims every day (10 TB of data/day) was ad hoc, heavily rule-based, and limited by data silos and a fragmented data environment.

Solution

UHG has developed a converged secure, enterprise platform for high-scale storage, NoSQL, and analytics. Their predictive analytics capabilities reduce medical costs by identifying mispaid claims in a systemic, repeatable manner, and they have developed and operate multiple predictive models. They enhance the data to get better structure, meta-data layers, graph analytics, and new data sources.

Benefits

- **Cost savings**
  - They save $250M/year by increasing claims efficiency.
  - They get a $22 return for every big data $1 spent.
- **Flexible platform** - Can integrate any new tool or technology seamlessly
- **Enterprise-grade features** - High availability and disaster recovery

- **Multi-tenancy** - Supports multiple business groups and applications in one cluster
- **Direct Access NFS** - Direct data ingestion, familiar access methods, existing tools work

MapR Risk Management Quick Start Solution

One of MapR's largest customer segments is the Financial Services industry. For many of the sub-segments of this industry, risk management concerns pervade their entire business.

The Risk Management Quick Start Solution from MapR is a data science-led product and services offering that addresses two major categories of risk for financial services companies:

- Fraud detection through predictive analytics
- Anti-money laundering through anomaly detection

This solution was designed to help financial services customer develop a systematic approach to fraud detection and anti-money laundering. This provides some key benefits:

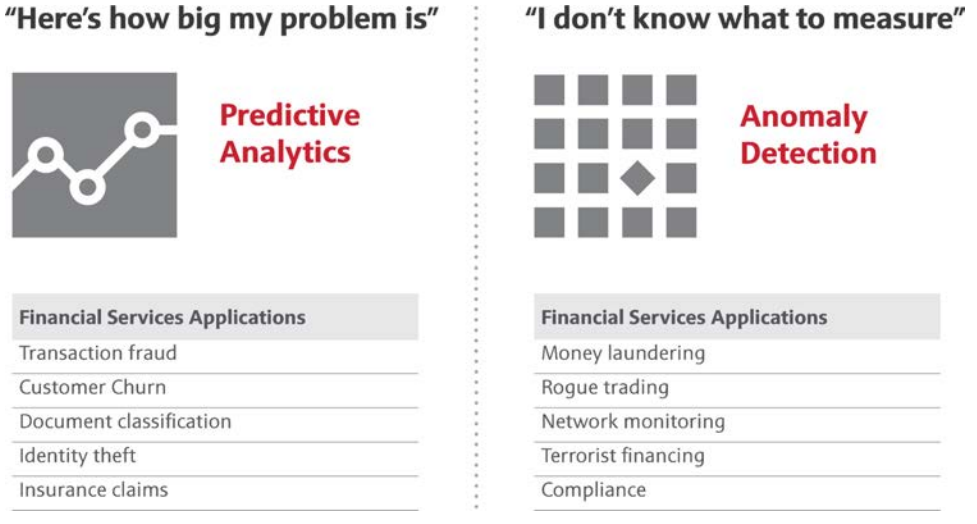
A scientific approach to risk: Customers engage with experienced MapR data scientists with financial services industry backgrounds.

Customized risk detection: MapR data scientists tailor the data models and algorithms for fraud detection and money laundering based on a collaborative process with a customer's fraud experts.

Precise ROI: A precise demonstration of the incremental monetary value to the business is provided as well as clear identification of suspicious activity, all delivered with minimal business disruption.

From a data science point of view, it makes sense to provide two vectors from which to tackle fraud and money laundering. Most fraud solutions will involve predictive analytics based upon fraud models that match the customer's sales model. These models are based on the premise that the customer has a clear definition of fraud and what that fraud is costing the business.

Money laundering is more difficult to identify and there is a real risk of false positives since there is no set pattern for how money laundering was accomplished. Data scientists therefore must use anomaly detection algorithms.



Staying with the data scientist for a moment, note from the above diagram the predictive analytics can be employed to identify customer churn, identity theft, and so on. Other anomaly detection use cases include rogue trading, terrorist financing. and regulatory compliance.

Operations

Referring to **the earlier discussion of "knowing versus doing,"** bringing big data technologies to operational systems is decidedly on the "doing" side of the equation. Some good examples of data-driven operational systems include:

- Assembly line yield optimization in manufacturing
- Network infrastructure monitoring and optimization for telcos
- Supply chain analytics and optimization for retail and Consumer Packaged Goods (CPG) companies
- Fleet optimization for logistics companies
- Predictive and preventative maintenance for oil and gas heavy machinery

Checklist to Progress to Phase IV: Optimization

Here are areas you should be thinking about as you progress to phase 4.

Organization

- Do you have commitments from all relevant lines of business in terms of executive sponsorship, program management, budget, and dedicated developers?
- Have you identified a Chief Data Officer, Chief Digital Officer, and business and technology leadership to make the digital transformation the new normal?
- Have you created clear roles and responsibilities within all lines of business for analytics, data curation, data lifecycle, and governance and data monetization?
- Have you communicated the new practices and capabilities that have resulted from your digital transformation and your data-driven approach?

Operations

- Have you developed a repeatable process to periodically re-evaluate operating parameters based on new technology and data-driven business practices?
- Have you created business and accounting processes that monetize data assets, track new revenue streams, and create measurable business service levels that extend beyond data center SLAs?
- Are you actively creating and managing a pipeline of new use cases to be deployed and supported?
- Are you managing capacity with optimal efficiency and do you have the metrics to prove it?
- Do you have security, privacy, governance, regulatory standards, and processes implemented for the new platform?
- Do you have reporting and measurement of service levels that are meaningful given your aspirational standards and practices?

Development and BI

- Do you you have dev, BI, and data science “rainmakers” that inject expertise, vitality, and innovation into your development and deployment?
- Do you have a comprehensive training regimen for developers and operations staff to continue refining digital processes and the technology that drive them?
- Has your portfolio of big data tools and technologies been thoroughly proven in production environments even during extreme conditions?
- Have you secured resources (internal or external) that readily provide expertise and/or staff augmentation in data science, data engineering, data visualizatoin, application architecture, and use case discovery?

Summary

Phase III: Expansion	
Description	Expand to multiple use cases across key lines of business
Motivation	Meet IT SLAs Line of Business competitiveness Create enterprise-grade operations, security, and governance **Accomplish ROI goals:** cost savings, efficiencies, and revenue generation Prepare for the post-big-data world Data Engineering Data Science
Executive Sponsor	**Multiple of:** CIO CTO VP of Development/Enterprise Applications VP of DW/Analytics LOB Executive CMO CISO CRO
Staffing	**Dedicated:** Cluster administrators (2-5) Developers (2-10) BI analysts (2-10) Dev ops (1) Data engineers (1-5) Data scientists (1-10) **Optional:** Chief Data Officer (CDO) Chief Analytics Officer
Participating Organizations/Groups	Central IT (infr) Application Development BI/Analytics Lines of Business (2-5) Marketing Sales Security Operations Finance
Hardware Investment	Dedicated clusters, distribute across data centers Public Cloud Hybrid Cloud
Number of Nodes	6 - 100

Key Capabilities	High-throughput, frictionless ingest Streaming ingest Enterprise-grade persistence/storage High Availability Replication Failover Mirroring Multi-tenancy Security, Privacy, and Governance Authorization Access Control Auditing Data protection Job/Data Placement Resource monitoring and management
Key Technologies	NFS MapR-FS MapR-DB MapR Streams Spark Spark Streaming Streaming Analytics Apache Drill
Key Skills Required	**IT operations:** Data management Data lifecycle Advanced data integration, cleansing Master data management Converged data ingest Hadoop administration Offload/extend life of legacy Storage offload (temperature-tiering) Ad hoc data engineering In-memory processing **Software development:** **Batch:** Hadoop Ecosystem **Interactive:** Spark **Query:** SQL on Hadoop **Search:** Solr, ElasticSearch Microservices development Streaming architectures **DWI/BI/Analytics team:** Predictive/preventative analytics NoSQL and real-time analytics Log analytics Schema-less ad hoc query (Drill, et al) Expand statistical skills (R)
# of Production Use Cases	5 - 20

Common Use Cases *from previous phase	Legacy offload: Data warehouse Mainframe Storage (NAS/SAN) IT focused: User home directories Machine log analyics File management Cold data offload/archive Data lake/hub: Analytics Platform Data Platform Vertical Data Lake (expert system) Analytics as a Service Application replatforming: Date Warehouse retirement RDBMS application re-engineering Reengineer legacy apps Marketing/Sales: Recommendation Engine Customer/patient/citizen 360 Next likely purchase Customer churn Ad/content/customer targeting Social/sentiment analysis Security: Security log analytics Fraud detection/prevention Advanced threat/intrusion detection SIEM system Operations: IoT/Industrial Internet Supply Chain optimization/analytics Logistics Predictive/preventative maintenance **Vertical specific** **Finance:** Trading Systems Risk Management Data Vault Market/Trading Analytics **Telecoms:** System-wide cost takeout Network monitoring, optimization Subscriber analytics Content/ad targeting Revenue management **Healthcare and life sciences:** Clinical decision support

	Fraud, waste and abuse Re-admission avoidance Smart devices and real-time patient monitoring Genomics **Retail and CPG:** Path to purchase Customer experience Ad targeting In-store operations Market basket analysis Pricing optimization
Number of Data Sources	15-100s
Data Sources	IT systems Date Warehouses/RDBMS File Servers, SAN/NAS Application/System Logs Clickstream ERP/CRM, SCM, HCM systems Billing systems External data sets Public data sources Data brokers

Phase 4 Optimization

Operate Shared Services and Deploy Converged Application Architectures

In this final phase, there is critical mass in the enterprise and all key LOBs are seeing the benefits of big data and advanced analytics. The data-driven processes that make up the digital business are now the new normal. You should leverage this capability to improve business and operational processes, which will reshape the business and give you a competitive edge. The goal is to be able to be predictive about most aspects of the business, and be able to respond and change operations in real time across more than one line of business.

However, companies that are operating at this level are disrupting their industries and demonstrating how responsive and adaptable a modern enterprise can be. Certain industries, such as ad tech, multinational finance, logistics, and social networks simply cannot exist without a complete commitment to digital transformation and operational agility. Like any important discipline, it is important to observe the most advanced players to help you up your own game, and to successfully progress through the four phases of big data adoption.

Disclaimer

The number of organizations at this advanced level is relatively small. Of all MapR customers, there are probably less than a dozen, including American Express, UnitedHealthcare Group, Ericsson, comScore, a Fortune 50 retailer, a Fortune 100 telecommunications company, and a few others. In the broader business landscape, one can cite many examples of companies that operate at this level: Google, LinkedIn, Facebook, and other born-of-the-web companies. But even for these companies, the new normal is still new. Therefore, it is difficult to be too prescriptive about what the new normal will look like.

Additionally, the most advanced practitioners understandably see their digital transformation as important intellectual property and a competitive advantage.

Motivating Factors

Meeting Enterprise Service Levels and Business Targets

As the digital business becomes the new normal, expectations for service levels rise in both an IT and a business context. IT systems are now expected to deliver essentially 100% availability and be able to routinely survive significant server, storage, network, and data center failures. The clustered architecture of the MapR Converged Data Platform provides multiple levels of redundancy and HA capabilities, and is a single security and governance entity that can greatly streamline business processes and regulatory compliance.

Achieving Operational Agility from Real-time Analytics

A data-driven business is looking to make real-time business decisions based on just in time information. But companies are looking for more than good decisions — they are looking for information and the ability to act on it. For manufacturers, it may mean creating an emergency maintenance window before a disastrous malfunction occurs. Digital processes in healthcare can anticipate problems with a patient based on real-time data from medical devices. Most industries have good cause to be able to react quickly to changing conditions on the ground.

Establish Advanced Governance and Development Methodology

One of the critical artifacts of today's IT operations is the data silos that have been created by "hand-rolling" different patchworks of systems/OSS projects. These silos slow down the data-to-action cycle, because data has to move between these different systems in the data pipeline. This is one of the basic motivators to begin a digital transformation in the first place. Different administration controls, security frameworks and data center issues with floor space, power, cooling, etc., are nightmare for IT operations to manage. Because of this, the big data program that powers your digital transformation provides a real opportunity to lay the groundwork for a new IT operations and software development methodology.

Key Activities and Use Cases

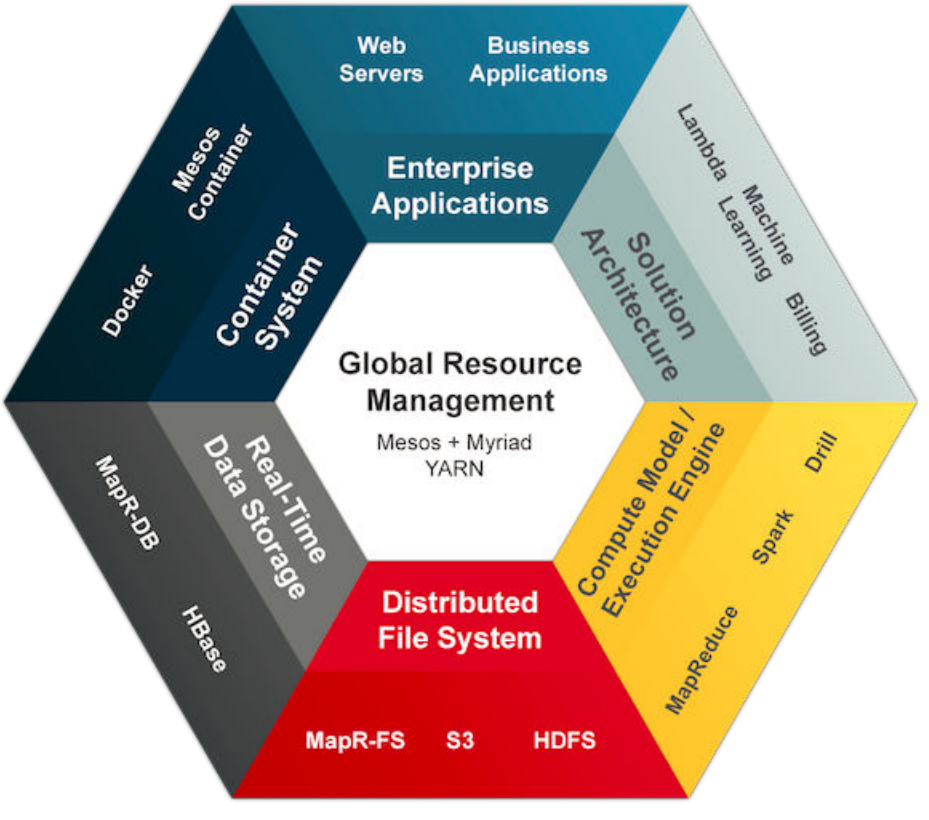
Established executive and organizational accountability

At this point, there is likely a designated Chief Data Officer (CDO) role, along with hired data scientists and data engineers that are looking for problems no one has solved yet. There is a deep investment and commitment to data science across all lines of business. Any actionable insights that are created have the potential to be automated. All key decision makers should have real-time visibility into the performance of their operations.

Formalize strategic architecture

Data-centric organizations such as Google and LinkedIn have pioneered architectures to deliver data and computation across thousands of servers, multiple data centers, and different geographies. All consumers now expect 100% uptime, instant answers, and personalized service. Why should enterprise data systems be any different?

Jim Scott, Director of Enterprise Strategy and Architecture at MapR, developed the Zeta Architecture, which is a high-level enterprise architectural construct not unlike the Lambda architecture which enables simplified business processes and defines a scalable way to increase the speed of integrating data into the business. The result? A powerful, data-centric enterprise.



There are seven pluggable components of the Zeta Architecture which work together, reducing system-level complexity while radically increasing resource utilization and efficiency.

Distributed file system - all applications read and write to a common, scalable solution, which dramatically simplifies the system architecture.

Real-time data storage - supports the need for high-speed business applications through the use of real-time databases.

Pluggable compute model / execution engine - delivers different processing engines and models in order to meet the needs of diverse business applications and users in an organization.

Deployment / Container management system - provides a standardized approach for deploying software. All resource consumers are isolated and deployed in a standard way.

Solution architecture - focuses on solving specific business problems, and combines one or more applications built to deliver the complete solution. These solution architectures encompass a higher-level interaction among common algorithms or libraries, software components, and business workflows.

Enterprise applications - brings simplicity and reusability by delivering the components necessary to realize all of the business goals defined for an application.

Dynamic and Global resource management - allows dynamic allocation of resources so that you can accommodate whatever task is the most important for that day.

Further Information on Zeta Architecture:

- ***Zeta Architecture on MapR.com***
- *Zeta Architecture* ***whitepaper***
- *Jim Scott delivers the Zeta Architecture* ***Whiteboard Walkthrough video***
- *Zeta Architecture:* ***Hexagon is the new circle*** *blog by Jim Scott*

A Note on Phase IV Use Cases

While big data use cases will become increasingly sophisticated over time, companies in the optimization phase or “peak digital transformation” will not necessarily be using state of the art technologies for all use cases. If the Expansion phase featured the development of groupings of related use cases into suites, the most advanced practitioners work in terms of practices.

A marketing suite of use cases simply becomes the new normal of marketing. Security use cases are assembled into a Security Information and Event Management platform. Analytics begin to pervade all lines of business and deliver real-time intelligence across a broad range of end users. While the nature of individual use cases may not be appreciably

different from Phase III, the ease and speed that new use cases are developed and deployed has accelerated. More importantly, the nature of enterprise solutions begins to change.

Developing new converged application models

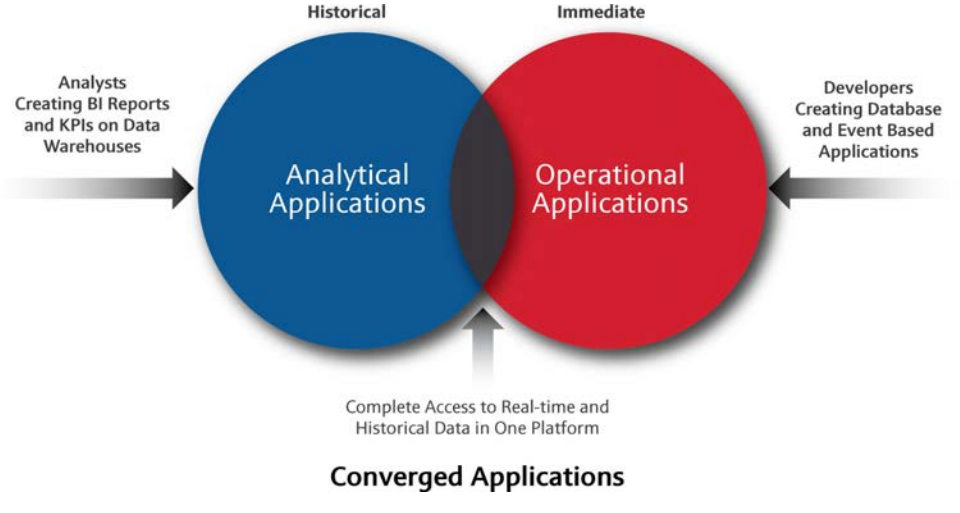
While it is likely that big data use cases will continue to evolve and be refined, it is also likely that application architectures and development methodologies will become a crucial part of that evolution.

Converged applications

Converged applications are software applications that can simultaneously process both operational and analytical data, allowing real-time, interactive access to both current and historical data. This class of applications deliver real-time analytics, high frequency decisioning, and other solution architectures that require immediate operations on large volumes of data.

Converged applications provide real-time access to large volumes of data in an efficient architecture to cost-effectively drive combined operational and analytical workloads on big data. They are often deployed in a modular architecture, especially as microservices that work together as a cohesive unit, not as monolithic processes in distinct data silos that require continual data movement. This architecture leads to greater responsiveness, better decisions, less complexity, lower cost, and lower risk.

Converged application architecture

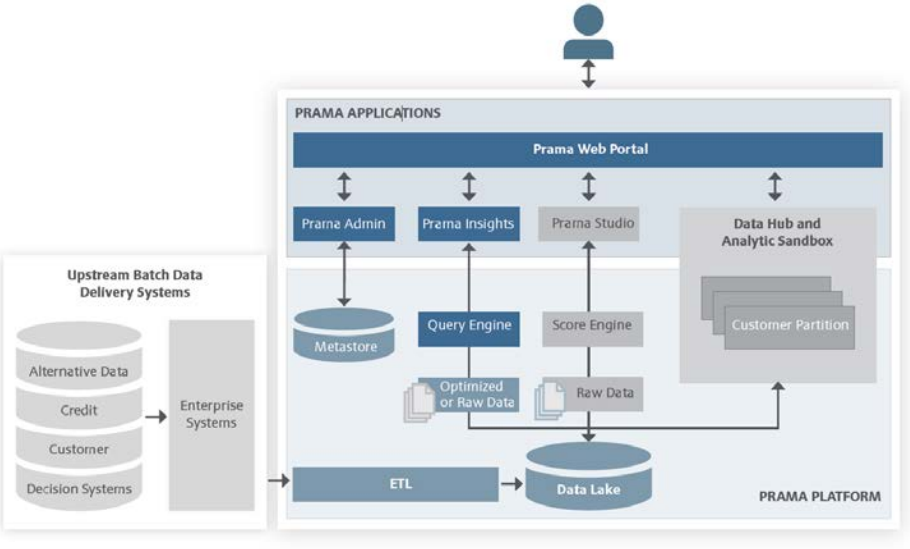


The key benefits of creating this new class of applications is the greater business value of immediate responses to events as they happen combined with the context provided by access to historical data. As the lines blur between operational and analytical systems, data movement is lessened, management overhead is reduced and therefore human error

and security gaps are minimized. By adopting this updated application model, you can future-proof your deployment because scaling up is a matter of simply adding more servers to the cluster.

A recent production example of a converged application architecture is the introduction of two new financial analytics solutions from TransUnion through their Prama platform. Prama Insights bases its analysis on TransUnion's anonymized consumer credit database and a seven-year historical view of data. Data sources include records compiled from over 85,000 data feeds, covering about 300 million consumers.

This self-service solution enables TransUnion customers to explore data and act on insights. With this new platform, TransUnion is allowing customers direct access to their content, but with the power of an advanced analytical platform and team of experts behind it.



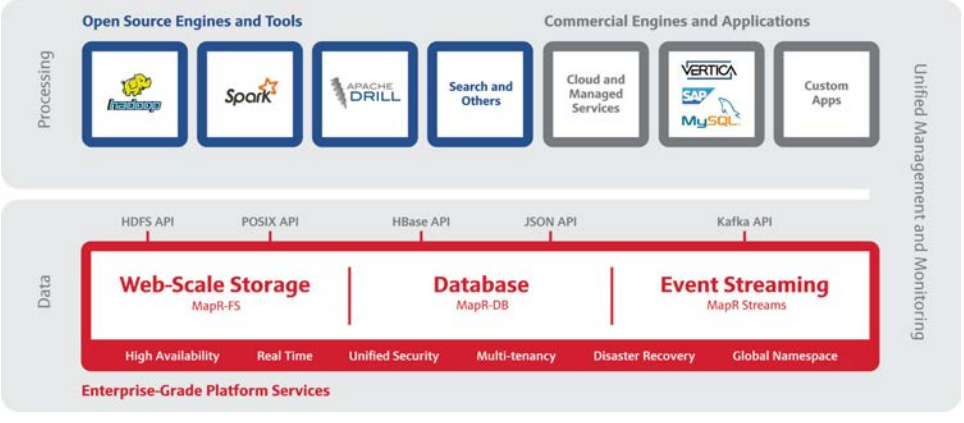
Looking at the underlying Prama architecture, it shows a mixture of batch delivery from their voluminous data feeds through an ETL process into a data lake. Prama then provides their customers with portal access into a personalized data hub and analytic sandbox.

Read more about the technology behind TransUnion Prama

The MapR Converged Data Platform in Digital Transformation

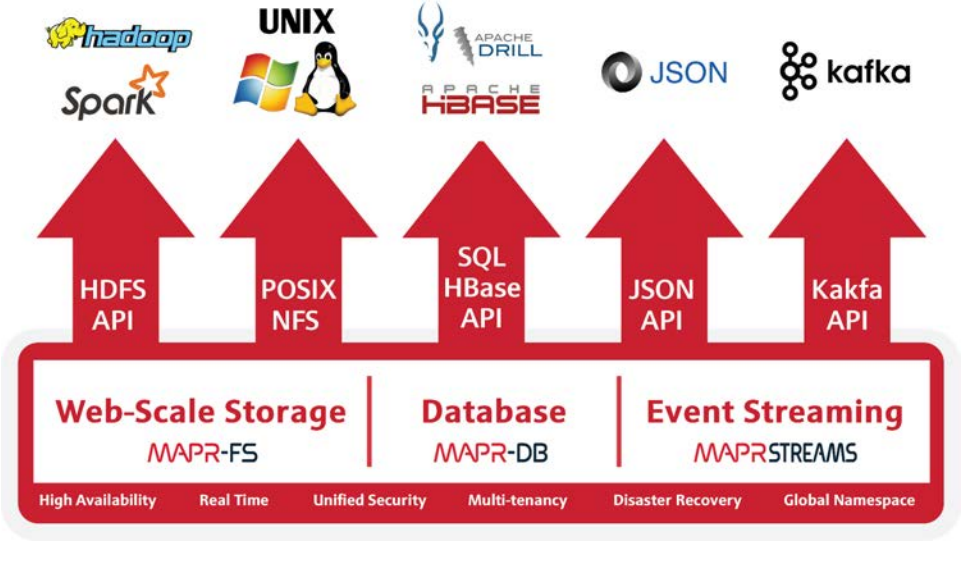
In 2014, Gartner introduced the concept of **Hybrid Transactional/Analytical Processing** or HTAP. They characterized this concept as a new set of systems and applications that could handle both traditional transaction processing workloads as well as OLAP-style ana-

lytical functions. The aforementioned Converged Application architecture and the MapR Converged Data Platform are designed to achieve the goals of HTAP and beyond.



Converged data management

The **Platform Services** of the MapR Converged Data Platform — **MapR-FS**, **MapR-DB**, and **MapR Streams** — provide core data management capabilities such as a global namespace, high availability, data protection, self-healing, unified security, real-time access, multi-tenancy, and management and monitoring.



A key design criteria of the MapR Platform is the strict use of existing enterprise standards and APIs. For easy data ingestion, MapR uses the NFS protocol and a POSIX standard file system, the same used on the vast majority of server systems in today's data centers. When practical, MapR has also relied on emerging open source standards and APIs such as Hbase, Apache Kafka APIs, and native support for JSON documents in MapR-DB.

It is the combination of open source projects merged with a hardened UNIX-style file system that encapsulates the fundamental strength of the MapR Platform.

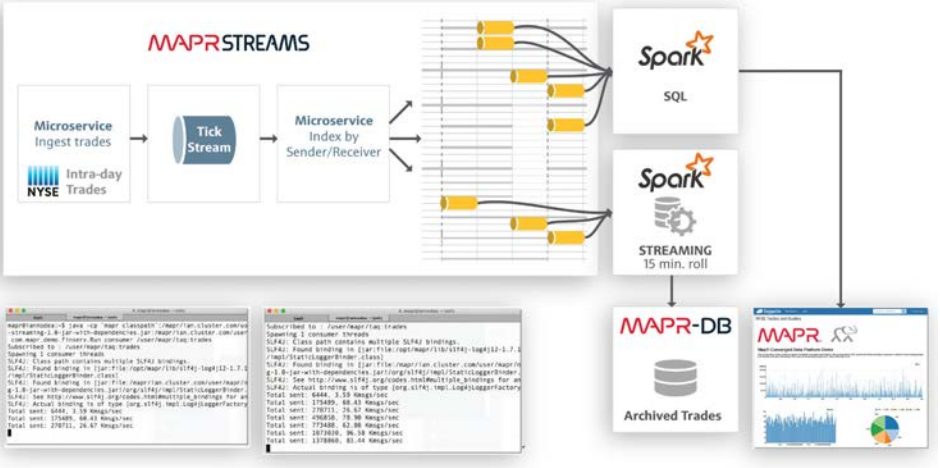
Streaming architectures and microservice development

In their book **Streaming Architecture: New Designs Using Apache Kafka and MapR Streams**, authors Ted Dunning and Ellen Friedman discuss the creation of new application architectures based upon microservices development and stream processing in order to deliver low latency analytic applications on a far larger scale. (More details about microservices can be found in **chapter 3 of "Streaming Architecture."**)

In a related Datanami article entitled **Streaming Architecture–Why Flow Instead of State?**, Dunning argues:

> Instead of a program with a finite input, we now have programs with infinite streams as inputs.... By adopting a streaming data architecture, we get a better fit between applications and real life. The advantages of this type of design are substantial: systems become simpler, more flexible and more robust. Multiple consumers can use the same streaming data for a variety of different purposes without interfering with each other. This independent multi-tenancy approach makes systems more productive and opens the way to data exploration that need not jeopardize existing processes. And where real time insights are needed, low latency analytics make it possible to react to life (and business) as it happens.

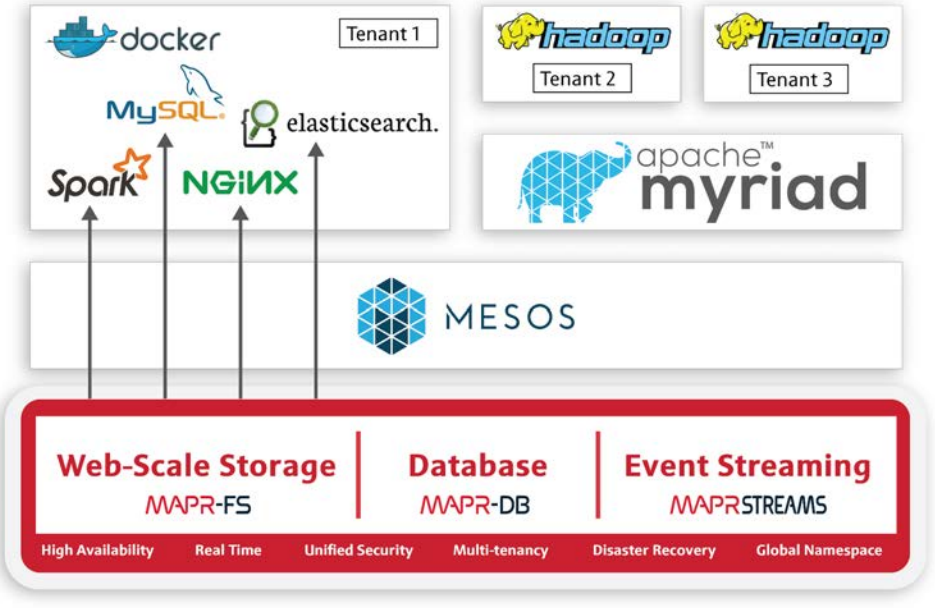
The following financial services converged application example illustrates the data pipeline and microservices of a financial services application that uses event streaming from MapR Streams, with microservices consuming and processing data from those streams, while Spark is used to query data in real time from an event stream or microservice output.



Containerization at the data platform layer

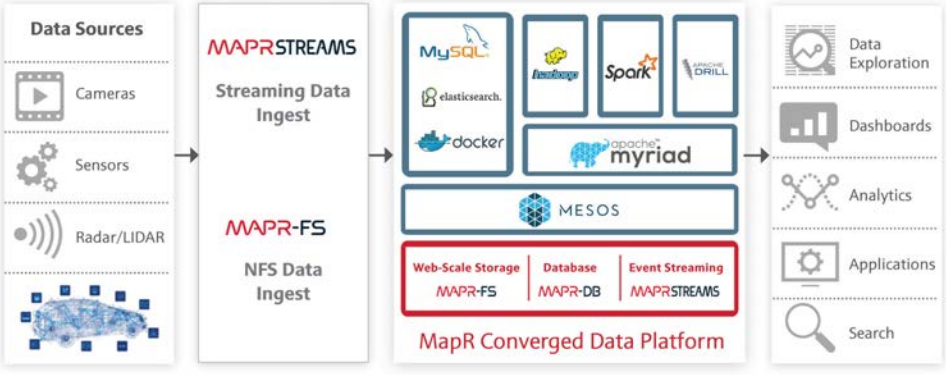
The advent of cloud computing has pushed IT architects to challenge how to push the limits of virtualization — and now containerization — to achieve operational agility, portability of applications and data, and rapid application/microservice provisioning. The big data world is following suit through technologies like Docker, Kubernetes and Apache Mesos.

Apache Myriad is a relatively new open source Hadoop project that lets YARN applications run side by side with Apache Mesos frameworks. It does this by registering YARN as a Mesos framework, and requesting Mesos resources on which to launch YARN applications. This allows YARN applications to run on top of a Mesos cluster without any modification.



Myriad is useful for organizations that use Hadoop with Docker and/or Apache Mesos and want to create a converged application environment between their enterprise applications and analytics. It lets you run Hadoop YARN applications on top of Apache Mesos clusters. This lets you share all resources, including data, across different workloads to improve time-to-value.

The combination of YARN, Docker, and Mesos makes up key components of the Zeta Architecture. The diagram below depicts a high-level deployment architecture that uses Mesos and Myriad in an automotive Internet of Things (IoT) context. While the Apache Myriad project currently resides with the Apache Incubator program, hopes are high that it will become an indispensable technology for cloud-based big data initiatives and beyond.



Further Information on Containerization:

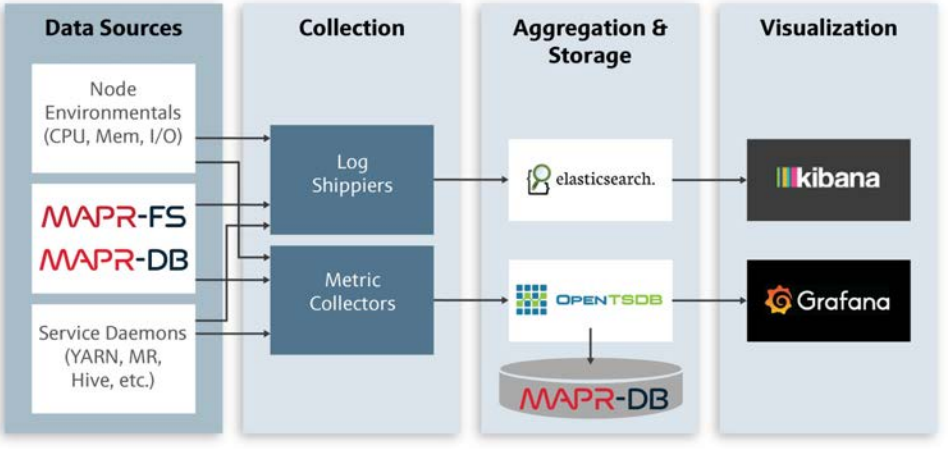
- **Apache Myriad on MapR.com**
- Apache Myriad on **incubator/apache.org**
- Fine-grained scaling with Apache Myriad **video**

Digital infrastructure monitoring: The Spyglass Initiative

The Spyglass Initiative is a multi-release MapR effort with the vision of increasing user and administrator productivity. It takes a comprehensive, open, and extensible approach to simplifying big data deployments.

Spyglass phase 1 – summer 2016: In the first phase of the Spyglass Initiative, MapR focuses on operational visibility to help customers with their ongoing big data successes. Successful big data deployments continue to get bigger and more complex. With new data sources, new use cases, new workloads, and new user groups, managing that growth requires a complete understanding of what is currently happening in the system.

MapR Monitoring helps you to manage successful big data deployments by giving you a converged, customizable, and extensible platform for cluster-wide visibility.



Conclusion

While it can be useful and instructive to look at maturity models as an indicator of your progress and a high level roadmap for future activities, there is a danger of getting hung up on which "bucket" you are currently in. The maturity model suggested in this document and the accompanying tables is somewhat subjective. Different organizations develop at different speeds along different axes, so it is important not to be overly concerned about where you fit on the maturity curve.

Some customers make amazing strides by successfully deploying a single use case. Others are focused on going deep in a particular line of business or application area such as supply chain, customer 360, or risk management. Others create a pipeline of dozens of use cases and quickmarch towards their digital transformation goals.

Whichever route you take, MapR will be there to support and inspire you to achieve great things with the MapR Converged Data Platform.

Summary

Phase IV: Optimization	
Description	Integrate and expand data-driven apps and analytics to all lines of business and more business functions
Motivation	Meet business SLAs 100% availability Redundancy

	Security Make real-time business decisions based on just-in-time information Eliminate data/technology silos Optimize operations Administrative controls Security frameworks Capacity and resource management
Executive Sponsor	**Many of:** CEO CFO COO CIO CTO CDO CRO CISO VP of DW/Analytics VP of Development/Enterprise Applications Multiple LOB Executives, GMs
Staffing	Chief Data Officer (CDO) Chief Analytics Officer Cluster administrators (2-5) Developers (5-100) BI analysts (5-50) Dev ops (1-3) Data engineers (10 - 30) Data scientists (5 - 100)
Participating Organizations/Groups	Central IT (infr) Application Development BI/Analytics Lines of Business (2-5) Marketing Sales Security Operations Finance
Hardware Investment	Large scale cluster, distributed across data centers Public Cloud Hybrid Cloud
Number of Nodes	50 - 1,000s
Key Capabilities	High-throughput, frictionless ingest Streaming ingest Enterprise-grade persistence/storage High Availability Replication Failover Mirroring Multi-tenancy Security, Privacy and Governance Authorization

	Access Control Auditing Data protection Job/Data Placement Resource monitoring and management
Key Technologies	NFS MapR-FS MapR-DB MapR Streams Spark Spark Streaming Streaming Analytics Apache Drill
Key Skills Required	**IT Operations:** Data Management Data Lifecycle Advanced data integration, cleansing Master data management Converged data ingest Hadoop administration Offload/extend life of legacy Storage offload (temperature-tiering) Ad hoc data engineering In-memory processing **Software Development:** **Batch:** Hadoop ecosystem **Interactive:** Spark, Query: SQL on Hadoop **Search:** Solr, ElasticSearch Microservices development Streaming architectures **DWI/BI/Analytics Team:** Predictive/preventative analytics NoSQL and real-time analytics Log analytics Schema-less ad hoc query (Drill, et al) Expand statistical skills (R)
# of Production Use Cases	20 - 100's
Common Use Cases *from previous phase	IT focused: User home directories Machine log analyics File management Cold data offload/archive Data lake/hub: Analytics platform Data platform Vertical data lake (expert system) Analytics as a service Application replatforming:

	Date warehouse retirement RDBMS application re-engineering Reengineer legacy apps Marketing/Sales: Recommendation engine Customer/Patient/Citizen 360 Next likely purchase Customer churn Ad/Content/Customer targeting Social/sentiment analysis Security: Security log analytics Fraud detection/prevention Advanced threat/intrusion detection SIEM system Operations: IoT/Industrial Internet Supply Chain optimization/analytics Logistics Predictive/preventative maintenance Vertical specific Finance: Trading systems Risk management Data vault Market/Trading analytics Telecoms: System-wide cost takeout Network monitoring, optimization Subscriber analytics Content/ad targeting Revenue management Healthcare and life sciences: Clinical decision support Fraud, waste and abuse Re-admission avoidance Smart devices and real-time patient monitoring Genomics Retail and CPG: Path to purchase Customer experience Ad targeting In-store operations Market basket analysis Pricing optimization
Number of Data Sources	100s-1000s
Data Sources	IT systems IoT and sensor networks Date warehouses File wervers, SAN/NAS Application/System logs

	Clickstream ERP/CRM, SCM, HCM systems External data sets Public data sources Data brokers

Maturity Model

Phase 1	Phase 2	Phase 3	Phase 4
Experimentation	**Implementation**	**Expansion**	**Optimization**
Understand capabilities of big data platform	*Develop first use cases and put into production*	*Expand to multiple use cases across key lines of business*	*Integrate and expand data driven apps and analytics to all lines of business and more business functions*

Motivation

Phase 1	Phase 2	Phase 3	Phase 4
Cost takeout IT preparedness Initiate big data strategy Develop staff skills Evaluate alternative technology	Creating a big data discipline Delivering measurable benefits Staff satisfaction and retention	Meet IT SLAs Line of Business competitiveness Create enterprise-grade operations, security, governance Accomplish ROI goals • Cost savings • Efficiencies • Revenue generation Prepare for the post big data world • Data Engineering • Data Science	Meet Business SLAs • 100% availability • Redundancy • Security Make real-time business decisions based on just-in-time information Eliminate data/technology silos Optimize operations • Administrative controls • Security frameworks • Capacity and resource management

Executive Sponsor

Phase 1	Phase 2	Phase 3	Phase 4
One of	**One or more of**	**Multiple of**	**Many of**
		CEO	CEO
		CFO	CFO
		COO	COO
		CDO	CDO
		CIO	CIO
CIO	CIO	CIO	CIO
CTO	CTO	CTO	CTO
Director of Development	VP of Development	VP of Development	VP of Development
Director of Analytics	VP Enterprise Applications	VP Enterprise Applications	VP Enterprise Applications
	VP of DW/Analytics	VP of DW/Analytics	VP of DW/Analytics
	LOB Executive	**LOB Executive**	**LOB Executive**
	CMO	CMO	CMO
	CRO	CRO	CRO
	CISO/CSO	CISO/CSO	CISO/CSO

Staffing

Phase 1	Phase 2	Phase 3	Phase 4
Part time	**Dedicated**	**Dedicated**	**Dedicated**
		Chief Data Officer (CDO) OPT*	Chief Data Officer (CDO)
		Chief Analytics Office OPT*	Chief Analytics Officer
Cluster administrator	Cluster administrator	Cluster administrators **2-5**	Cluster administrators **2-5**
Developers	Developers **1-3**	Developers **2-10**	Developers **5-100**
BI analysts	BI analysts	BI analysts **2-10**	BI analysts **5-50**
	Developer ops PT*	Developer ops **1**	Developer ops **1-3**
	Data engineer PT/*OPT*	Data engineers **1-5**	Data engineers **10-30**
	Data scientist OPT*	Data scientists **1-10**	Data scientists **5-100**
**PT = Part time* **OPT = Optional*	Staff augmentation (consultants/contractors)		

Participating Organizations/Groups

Phase 1	Phase 2	Phase 3	Phase 4
Central IT (infr)	Central IT (infr)	Central IT (infr)	Central IT (infr)
Application Development	Application Development	Application Development	Application Development
BI/Analytics	BI/Analytics	BI/Analytics	BI/Analytics
	Lines of Business	**Lines of Business**	**Lines of Business**
	• Marketing	• Marketing	• Marketing
	• Sales	• Sales	• Sales
	• Security	• Security	• Security
		• Operations	• Operations
		• Finance	• Finance

Hardware Investment

Phase 1	Phase 2	Phase 3	Phase 4
	Dedicated cluster	Dedicated clusters, distribute across data centers	Large scale cluster, distributed across data centers
Public Cloud	Public cloud	Public Cloud	Public Cloud
		Hybrid Cloud	Hybrid Cloud
Servers on Premises			
• Borrowed/shared			
• Near end of life			
• Off-the-books hardware			

Number of Nodes

Phase 1	Phase 2	Phase 3	Phase 4
2-5	3-6	6-100	50-1000s

Key Capabilities

Phase 1	Phase 2	Phase 3	Phase 4
	High-throughput, frictionless ingest	High-throughput, frictionless ingest	High-throughput, frictionless ingest
Easy ingest			
		Streaming ingest	Streaming ingest
Easy installation			
Application templates, scripts			
	Enterprise grade persistence/storage	Enterprise grade persistence/storage	Enterprise grade persistence/storage
	High Availability	High Availability	High Availability
	• Replication	• Replication	• Replication
	• Failover	• Failover	• Failover
	• Mirroring	• Mirroring	• Mirroring
	Multi-tenancy	Multi-tenancy	Multi-tenancy
	Basic security	Security, Privacy and Governance	Security, Privacy and Governance
	• Authorization	• Authorization	• Authorization
	• Access control	• Access control	• Access control
		• Auditing	• Auditing
		• Data protection	• Data protection
		Job/Data Placement	Job/Data Placement
		Resource monitoring and management	Resource monitoring and management

Key Technologies

Phase 1	Phase 2	Phase 3	Phase 4
Easy ingest (NFS)	NFS	NFS	NFS
File System	MapR-FS	MapR-FS	MapR-FS
NoSQL Database	MapR-DB	MapR-DB	MapR-DB
		MapR-Streams	MapR-Streams
	Apache Spark (core)	Spark	Spark
		Spark Streaming	Spark Streaming
		Streaming Analytics	Streaming Analytics
	Apache Drill	Apache Drill	Apache Drill

Key Skills Required

Phase 1	Phase 2	Phase 3	Phase 4
IT Operations	**IT Operations**	**IT Operations**	**IT Operations**
Basic software/cluster install and data management	Data management	Data Management	Data Management
Manual data ingest	Data ingest	Data Lifecycle	Data Lifecycle
		Advanced data integration, cleansing	Advanced data integration, cleansing
		Master data management	Master data management
		Converged data ingest	Converged data ingest
Basic Hadoop admin	Hadoop administration	Hadoop administration	Hadoop administration
	Offload/extend life of legacy	Offload/extend life of legacy	Offload/extend life of legacy
	Storage offload	Storage offload (temperature-tiering)	Storage offload (temperature-tiering)
	Ad hoc data engineering	Ad hoc data engineering	Ad hoc data engineering
		In-memory processing	In-memory processing
Software development/BI	**Software development**	**Software development**	**Software development**
Batch: MapReduce	Batch: Hadoop Ecosystem	Batch: Hadoop Ecosystem	Batch: Hadoop Ecosystem
Interactive: SQL on Hadoop (Drill, Impala, Hive)	Interactive: Spark	Interactive: Spark	Interactive: Spark
	Query: SQL on Hadoop	Query: SQL on Hadoop	Query: SQL on Hadoop
	Search: Solr, Elastic	Search: Solr, Elastic	Search: Solr, Elastic
	Microservice pilots	Microservices development	Microservices development
		Streaming architectures	Streaming architectures
DWI/BI/Analytics Team	**DWI/BI/Analytics Team**	**DWI/BI/Analytics Team**	**DWI/BI/Analytics Team**
		Predictive/preventative analytics	Predictive/preventative analytics
Basic NoSQL	NoSQL	NoSQL and real-time analytics	NoSQL and real-time analytics
Log analytics	Log analytics	Log analytics	Log analytics
	Schema-less ad hoc query (Drill, et al)	Schema-less ad hoc query (Drill, et al)	Schema-less ad hoc query (Drill, et al)
Basic statistical skills (R)	Expand statistical skills (R)	Expand statistical skills (R)	Expand statistical skills (R)

Number of Production Use Cases

Phase 1	Phase 2	Phase 3	Phase 4
0-1	1-5	5-20	20-1000s

Common Use Cases

Phase 1	Phase 2	Phase 3	Phase 4
Platform Bakeoffs			
Hadoop			
Streaming			
NoSQL			
Analytics			
SQL on Hadoop			
Legacy Offload			
DW			
Mainframe			
Storage			
IT Focused			
User home directories			
Machine log analytics (trial)			
File management			
Cold data offload/archive			
Security log analytics			
	Application Replatforming		
	Date Warehouse takeout		
	RDBMS takeout		
	Re-engineered legacy apps		
	Data Lake/Hub	**Data Lake/Hub**	**Data Lake/Hub**
	Analytics Platform	Analytics Platform	Analytics Platform
	Data Platform	Data Platform	Data Platform
		Analytics as a Service	**Analytics as a Service**
	Marketing/Sales	**Marketing/Sales**	**Marketing/Sales**
	Customer 360	Customer 360	Customer 360
	Recommendation Engine	Recommendation Engine	Recommendation Engine
	Next likely purchase	Next likely purchase	Next likely purchase
		Ad/content targeting	Ad/content targeting
		Social/sentiment analysis	
	Security	**Security**	**Security**
	Security log analytics	Security log analytics	Security log analytics
	Fraud detection	Fraud detection	Fraud detection
		Advanced threat/intrusion detection	Advanced threat/intrusion detection
		Operations, vertical specific	**Operations, vertical specific**
		IoT	IoT
		Supply Chain and Logistics	Supply Chain and Logistics
		Assembly Line (QA and Yield Optimization)	Assembly Line (QA and Yield Optimization)
		Predictive/preventative maintenance	Predictive/preventative maintenance
			Specialty Vertical
			Telecoms
			• Network/bandwidth optimization
			• System-wide cost takeout
			• Subscriber analytics
			Healthcare
			• Clinical decision support
			• Patient data management
			• Fraud, waste and abuse prevention
			• Genomics
			Retail
			• Path to purchase
			• Market basket analysis
			• Price optimization
			Financial Services
			• Trading systems
			• Risk Management
			• Data vault

Number of Data Sources

Phase 1	Phase 2	Phase 3	Phase 4
1-5	5-15	15-100s	100s-1000s

Data Sources

Phase 1	Phase 2	Phase 3	Phase 4
Legacy servers and storage (files)	IT systems	IT systems	IT systems
Data Warehouses/RDBMS	Data Warehouses/RDBMS	Data Warehouses/RDBMS	Data Warehouses
File Servers, SAN/NAS	File Servers, SAN/NAS	File Servers, SAN/NAS	File Servers, SAN/NAS
Sample Data			
	Application/System Logs	Application/System Logs	Application/System Logs
	Clickstream	Clickstream	Clickstream
	ERP/CRM systems	ERP/CRM, SCM, HCM systems	ERP/CRM, SCM, HCM systems
		Billing systems	
		External data sets	External data sets
		• Public data sources	• Public data sources
		• Data brokers	• Data brokers
			IoT and sensor networks