(Big-)Data Architecture (Re-)Invented
Part 1: Hadoop and Data Lake

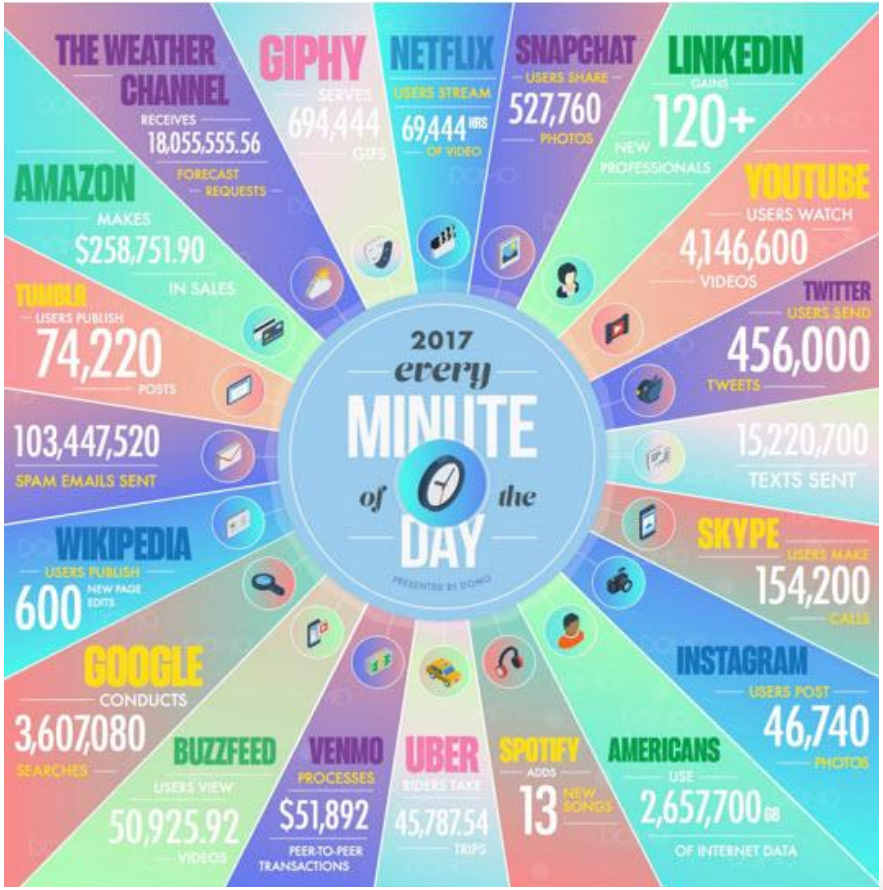
William El Kaim
May 2018 – V 4.0

This Presentation is part of the

Enterprise Architecture Digital Codex

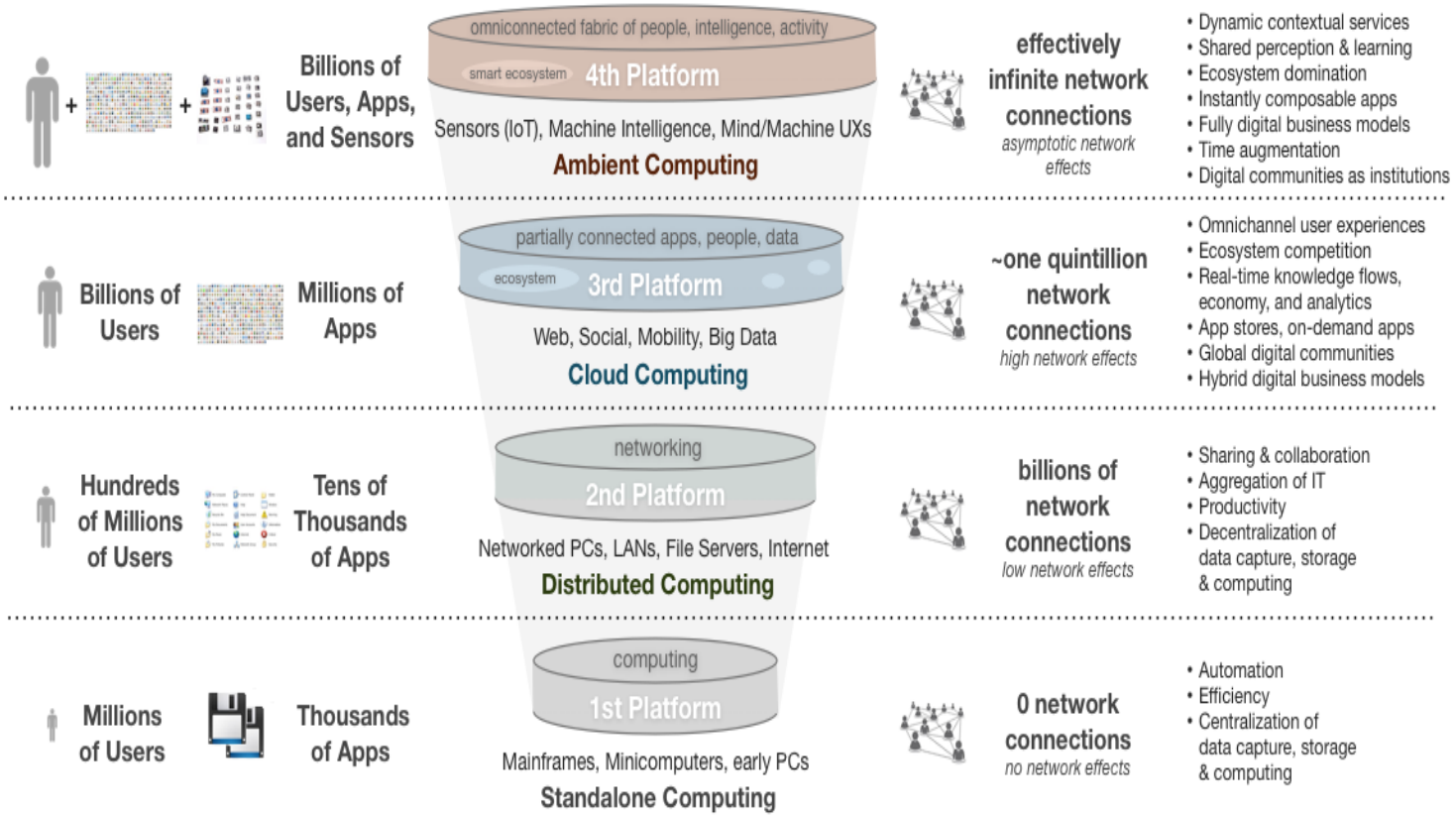
http://www.eacodex.com/

Taming the Data Deluge (2017)

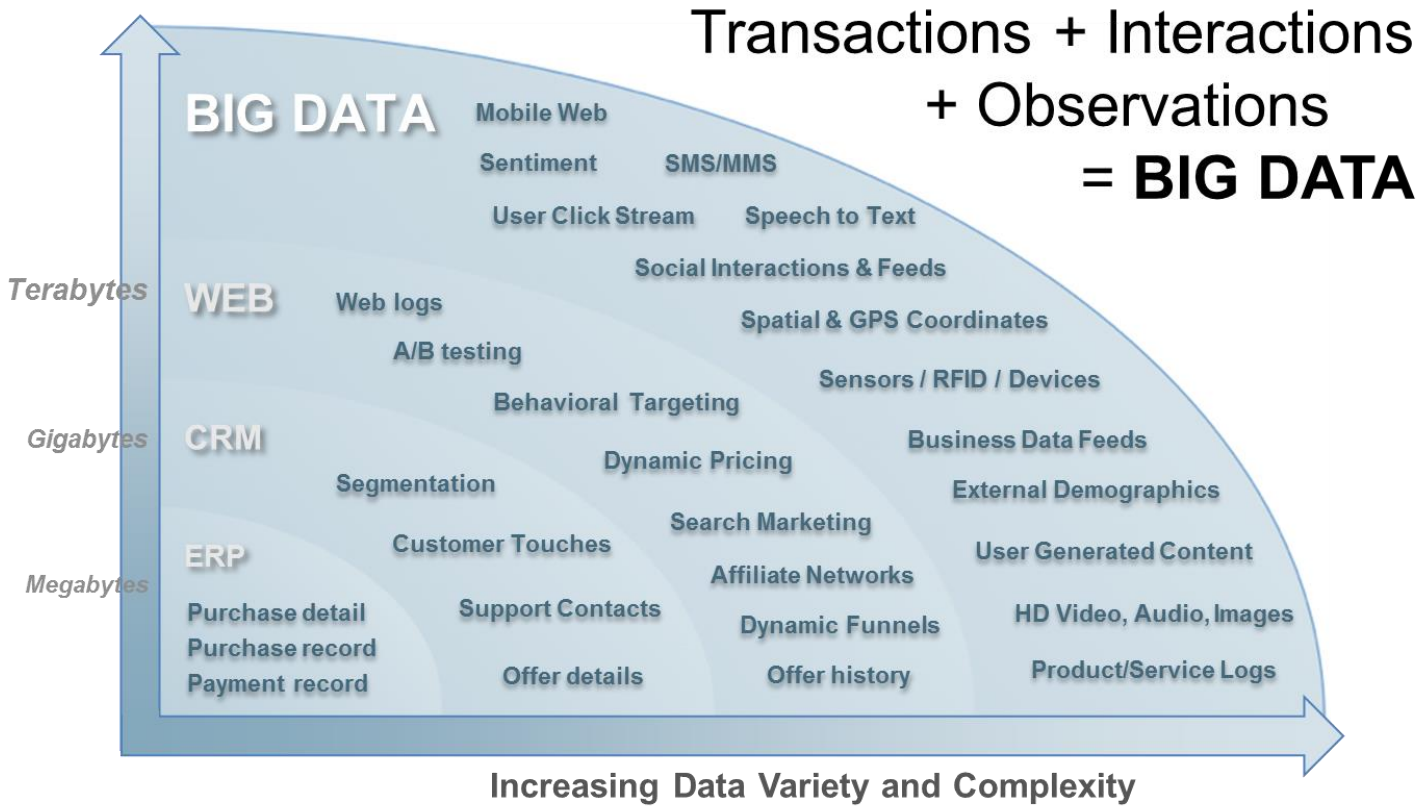


Source: LuceMedia

Taming the Data Deluge



Taming the Data Deluge



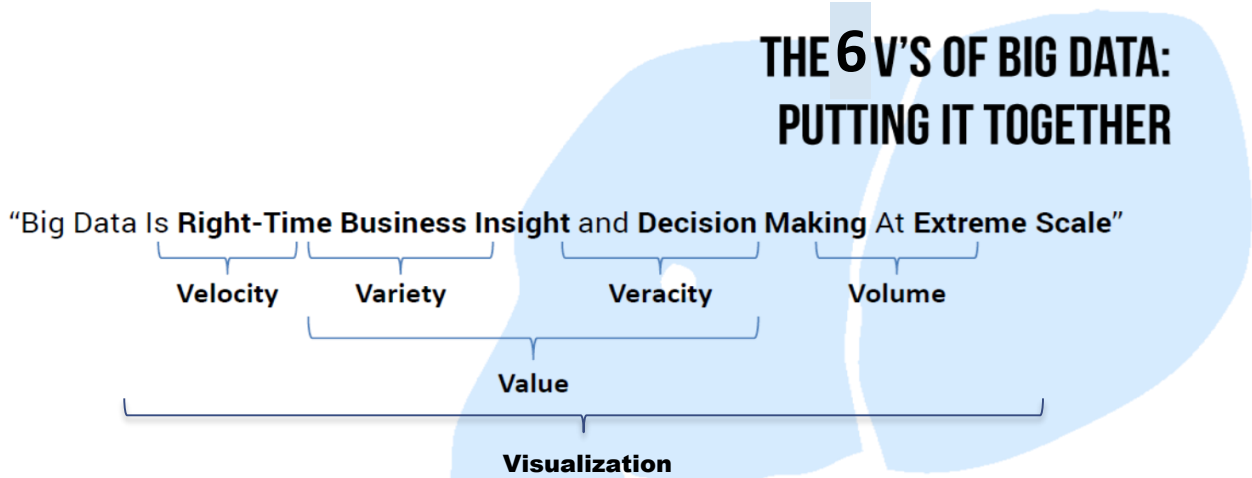
- Taming The Data Deluge
- What is Big Data?
- Why Now?
- What is Hadoop?
- What is Hadoop Data Lake?
- When to use Hadoop?
- Getting Started with Big Data

What is Big Data?

- A collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
  - Due to its technical nature, the same challenges arise in Analytics at much lower volumes than what is traditionally considered Big Data.
- Other definitions
  - When the data could not fit in Excel
    - Used to be 65,536 lines, Now 1,048,577 lines
  - When it's cheaper to keep everything than spend the effort to decide what to throw away (David Brower @dbrower)

Source: SiSense

What is Big Data?



Source: James Higginbotham

The “Vs” to Nirvana

With the datafication comes big data, which is often described using the four Vs: Volume, Velocity, Variety and Veracity

Volume refers to the vast amounts of data generated every second. We are not talking Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. New big data tools use distributed systems so that we can store and analyse data across databases that are dotted around anywhere in the world.

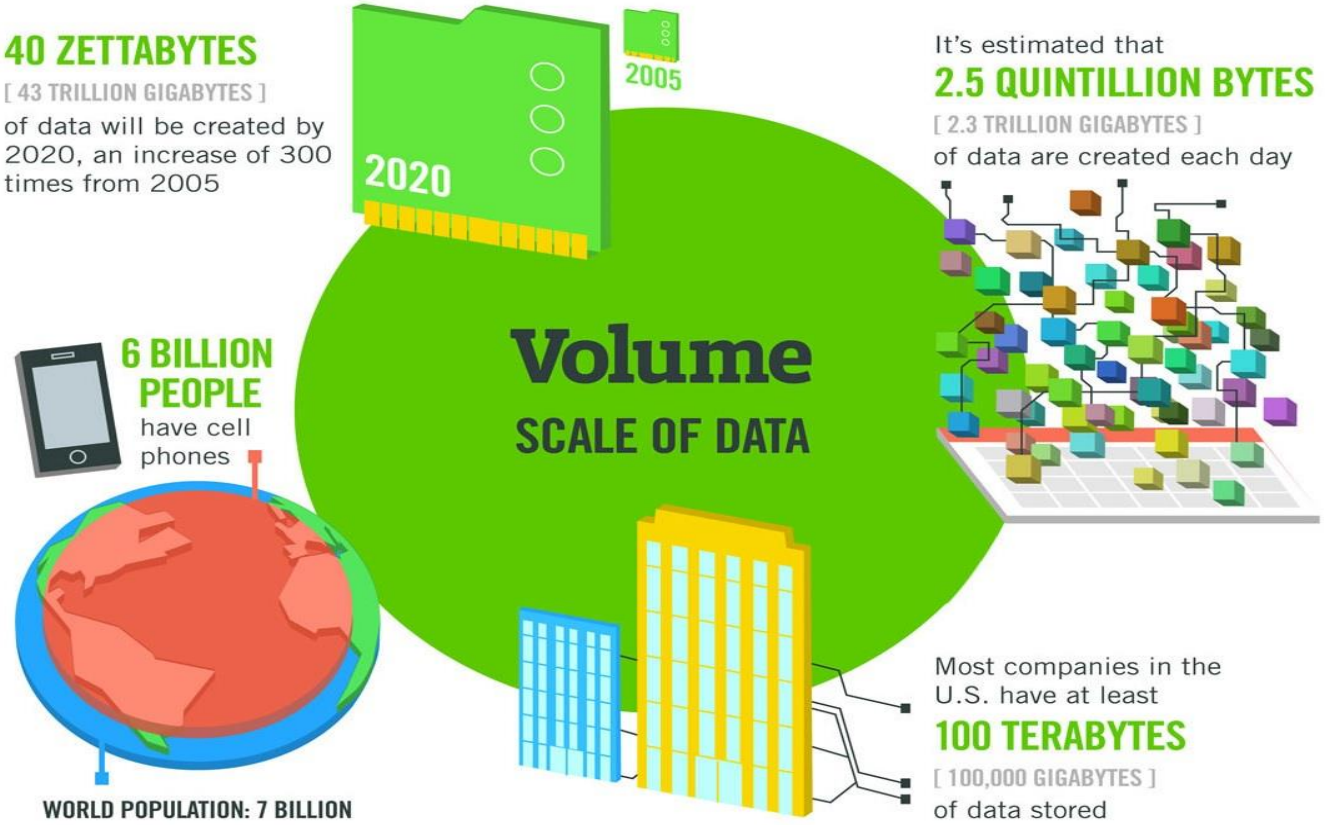
Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology allows us now to analyse the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

Variety refers to the different types of data we can now use. In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world’s data is unstructured (text, images, video, voice, etc.) With big data technology we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

Veracity refers to the messiness or trustworthiness of the data. With many forms of big data quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but technology now allows us to work with this type of data.

Source: Bernard Marr

The “Vs” to Nirvana



Source: IBM

The “Vs” to Nirvana

The New York Stock Exchange captures

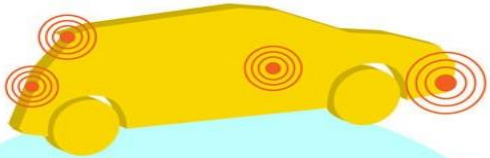
1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure



Velocity

ANALYSIS OF STREAMING DATA

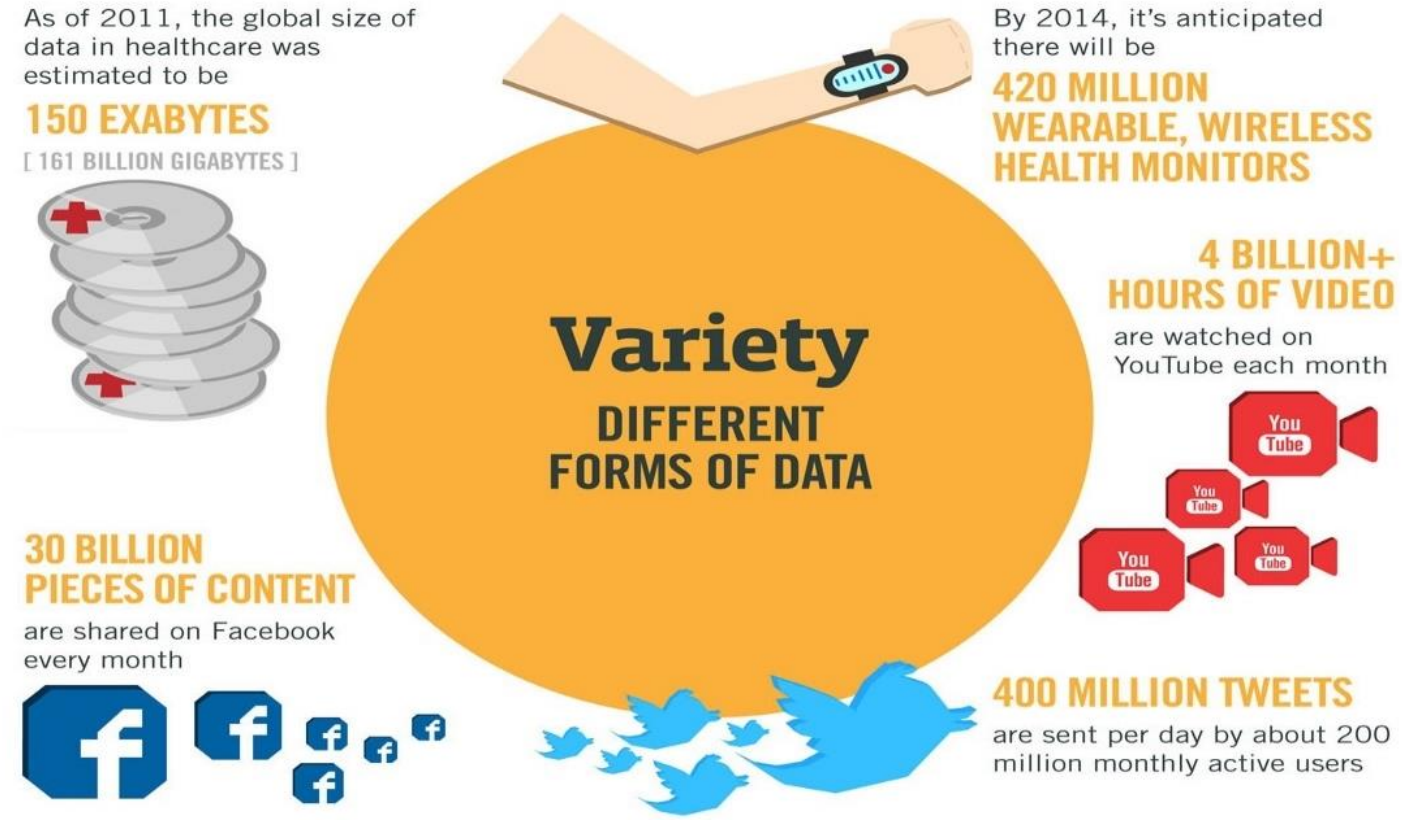
By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

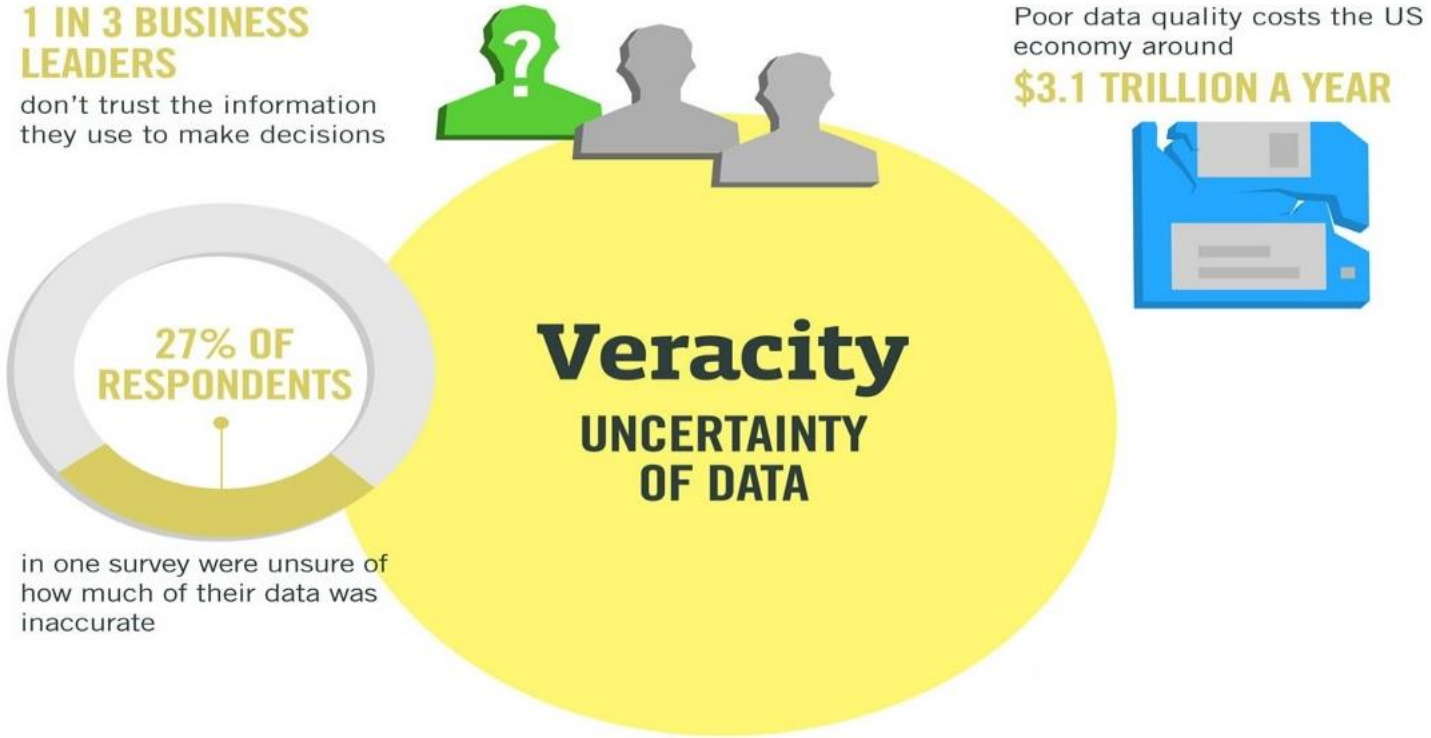
Source: IBM

The “Vs” to Nirvana



Source: IBM

The “Vs” to Nirvana



Source: IBM

The “Vs” to Nirvana

The ‘Datafication’ of our World;

- Activities
- Conversations
- Words
- Voice
- Social Media
- Browser logs
- Photos
- Videos
- Sensors
- Etc.

Volume

Velocity

Variety

Veracity

Analysis

Analysing Big Data:

- Text analytics
- Sentiment analysis
- Face recognition
- Voice analytics
- Movement analytics
- Etc.

Visualization

Value

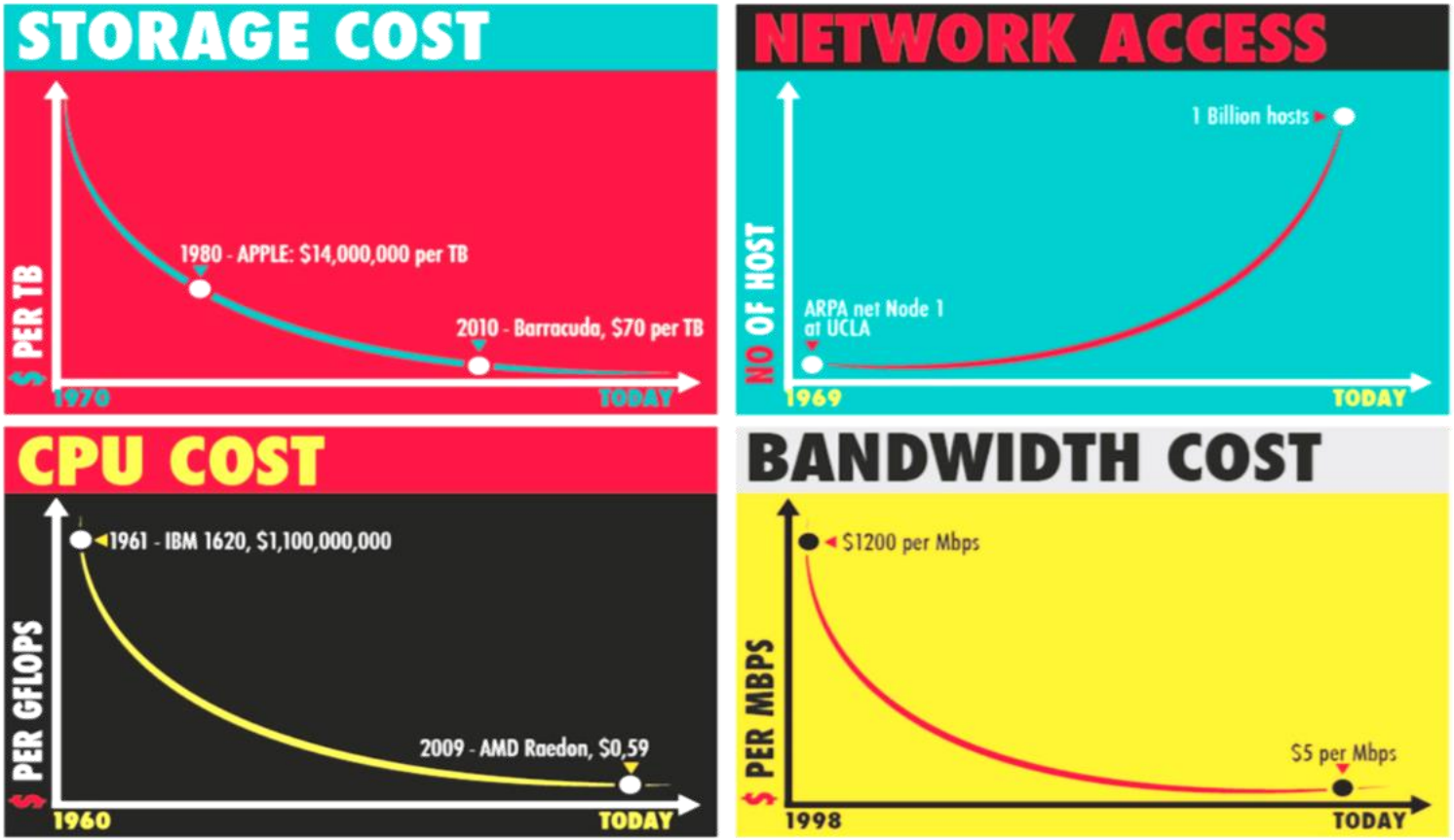
Source: Bernard Marr

The “Vs” to Nirvana

Source: M-Brain

“Big Data” Landscape

Why Now?

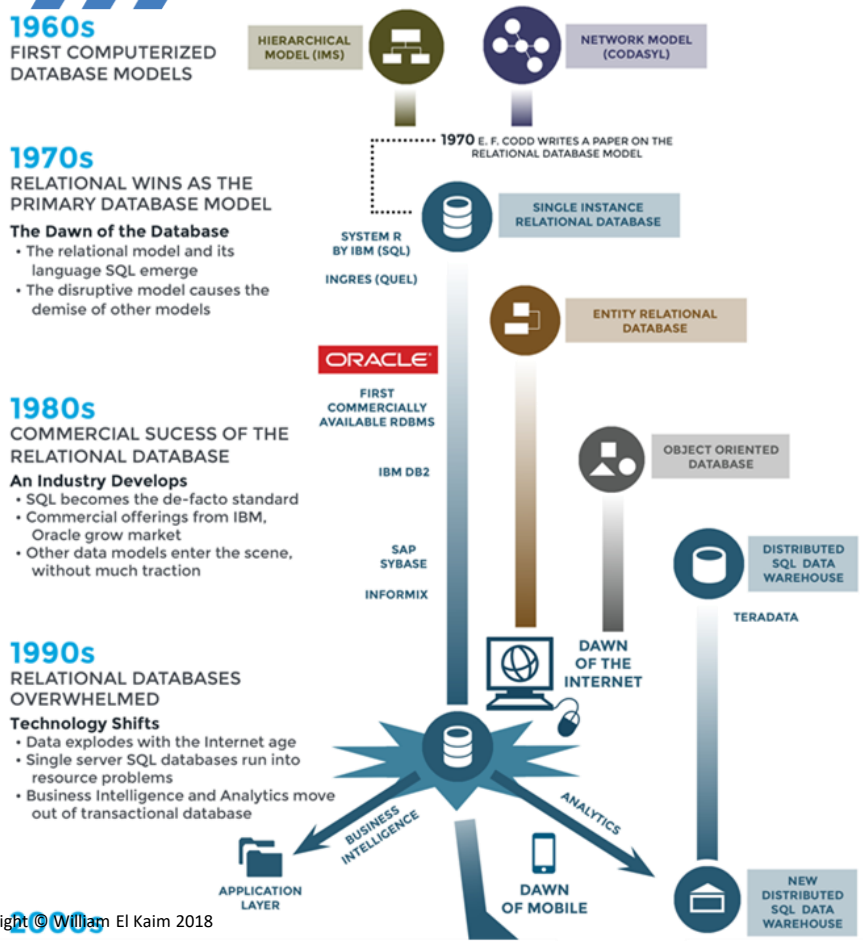


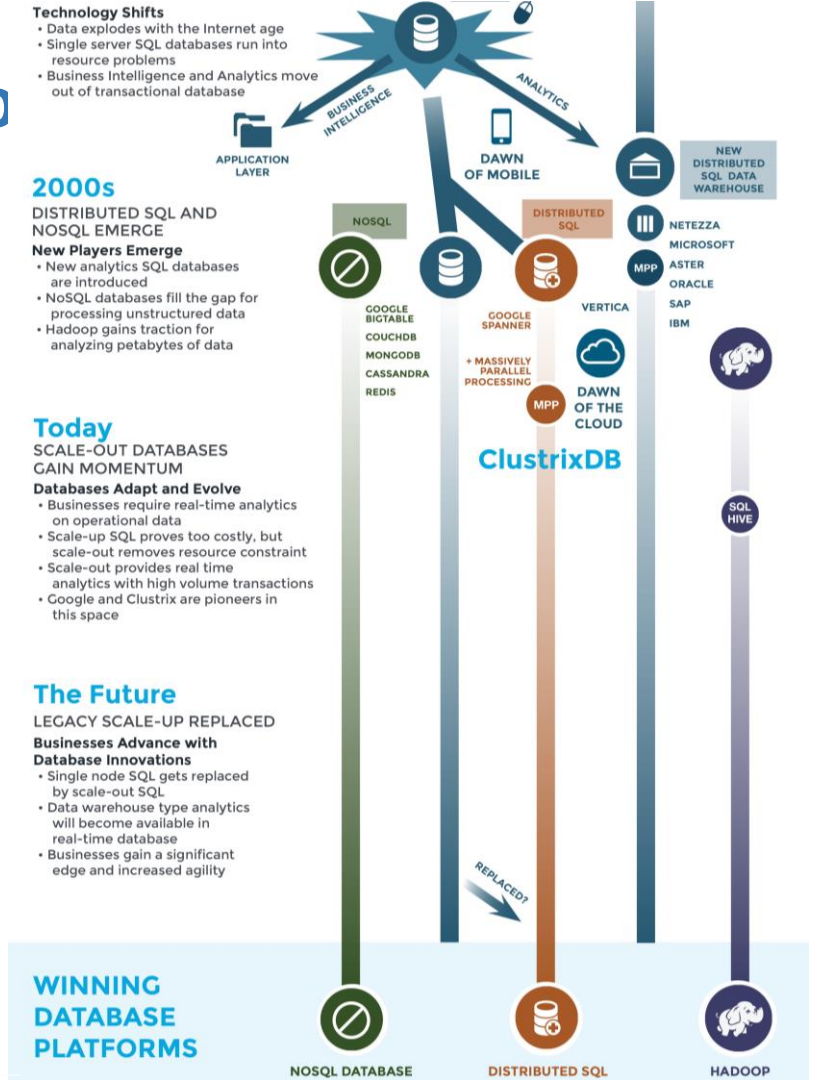
Credit to: Mike Driscoll, CTO Metamarkets: The Three Sexy Skills of Data Scientists (& Data Driven Startups)

Why Now? Datafication of the World

Source: Bernard Marr

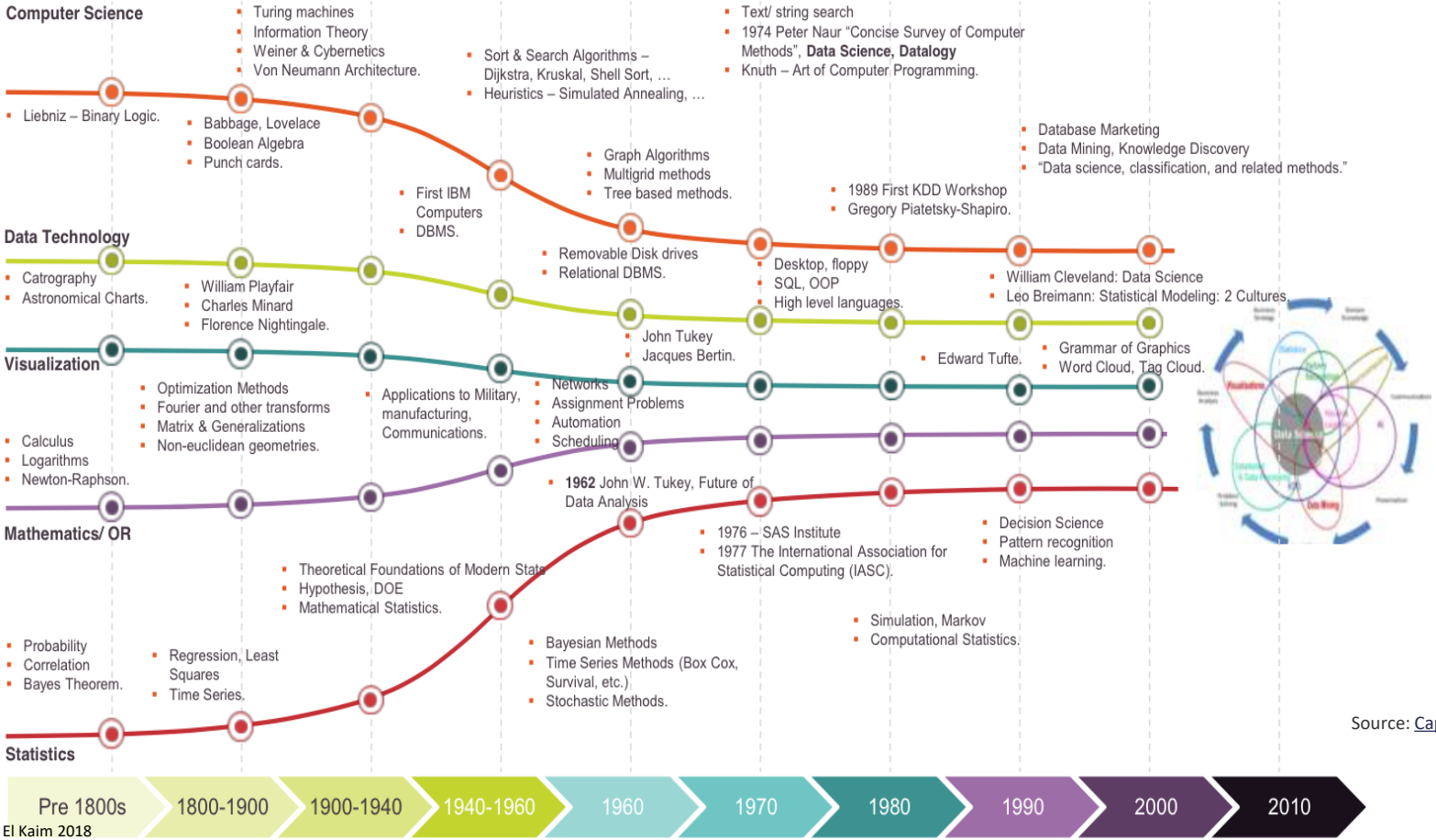
Databases Evolutio





Source: Clustrix

Why Now? Data Science (R)evolution



Source: Capgemini

Why Now? Open APIs, Platforms, Ecosystems

ebizQ From http://ebizq.net/blogs/enterprise by Dion Hinchcliffe

What is Hadoop?

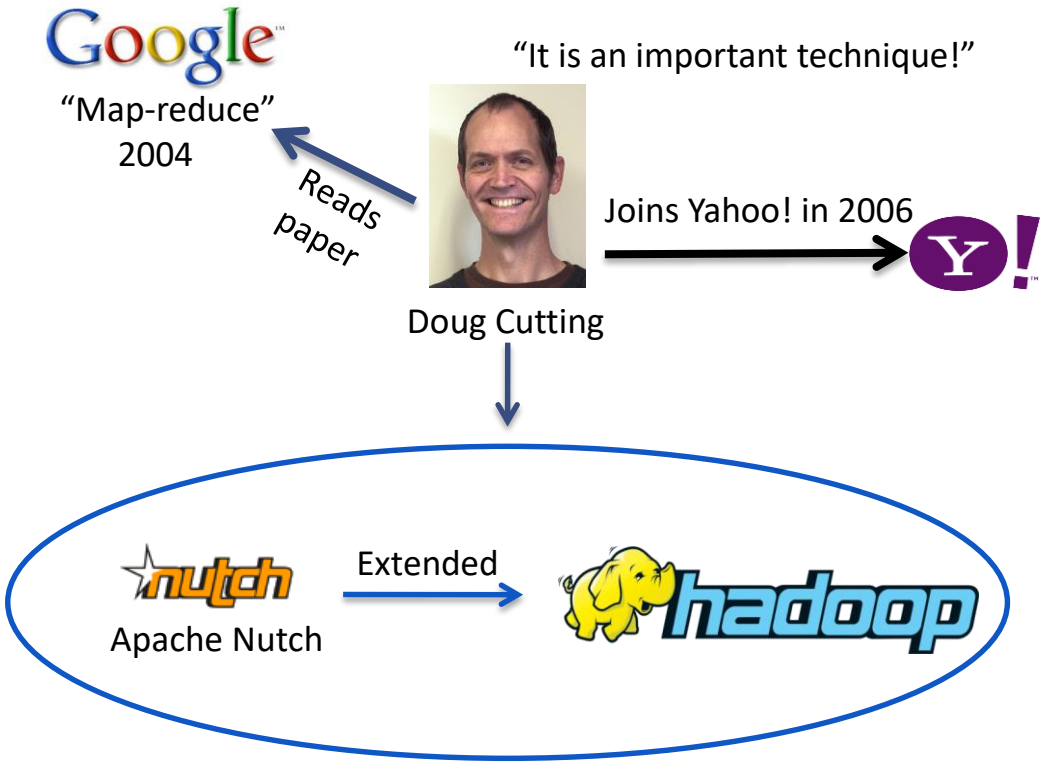
- Apache Hadoop is an open source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.
  - Hadoop services provide for data storage, data processing, data access, data governance, security, and operations.

Why Hadoop?

- Scalability issue when running jobs processing terabytes of data
  - Could span dozen days just to read that amount of data on 1 computer
- Need lots of cheap computers
  - To Fix speed problem
- But lead to reliability problems
  - In large clusters, computers fail every day
  - Cluster size is not fixed
- Need common infrastructure
  - Must be efficient and reliable

Hadoop Genesis

- Google File System paper published in October 2003.
- This paper spawned another research paper from Google
  - MapReduce: Simplified Data Processing on Large Clusters.
- Development started in the Apache Nutch project, but was moved to the new Hadoop subproject in January 2006.
- The first committer added to the Hadoop project was Owen O'Malley in March 2006.
- Hadoop 0.1.0 was released in April 2006.



Source: Xiaoxiao Shi

Hadoop Genesis

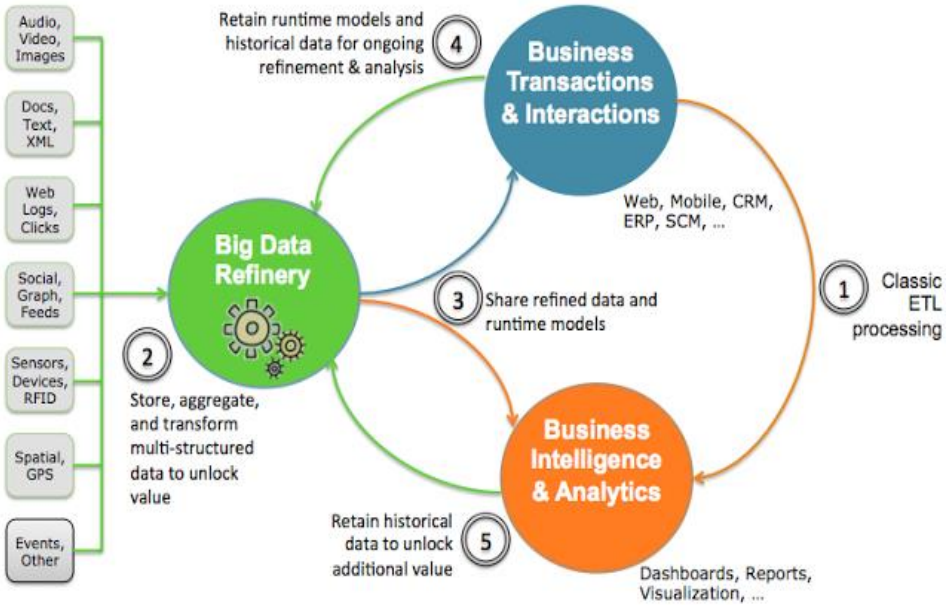
"Hadoop came from the name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria."

Doug Cutting, Hadoop project creator

- Open Source Apache Project
- Written in Java
- Running on Commodity hardware and all major OS
  - Linux, Mac OS/X, Windows, and Solaris

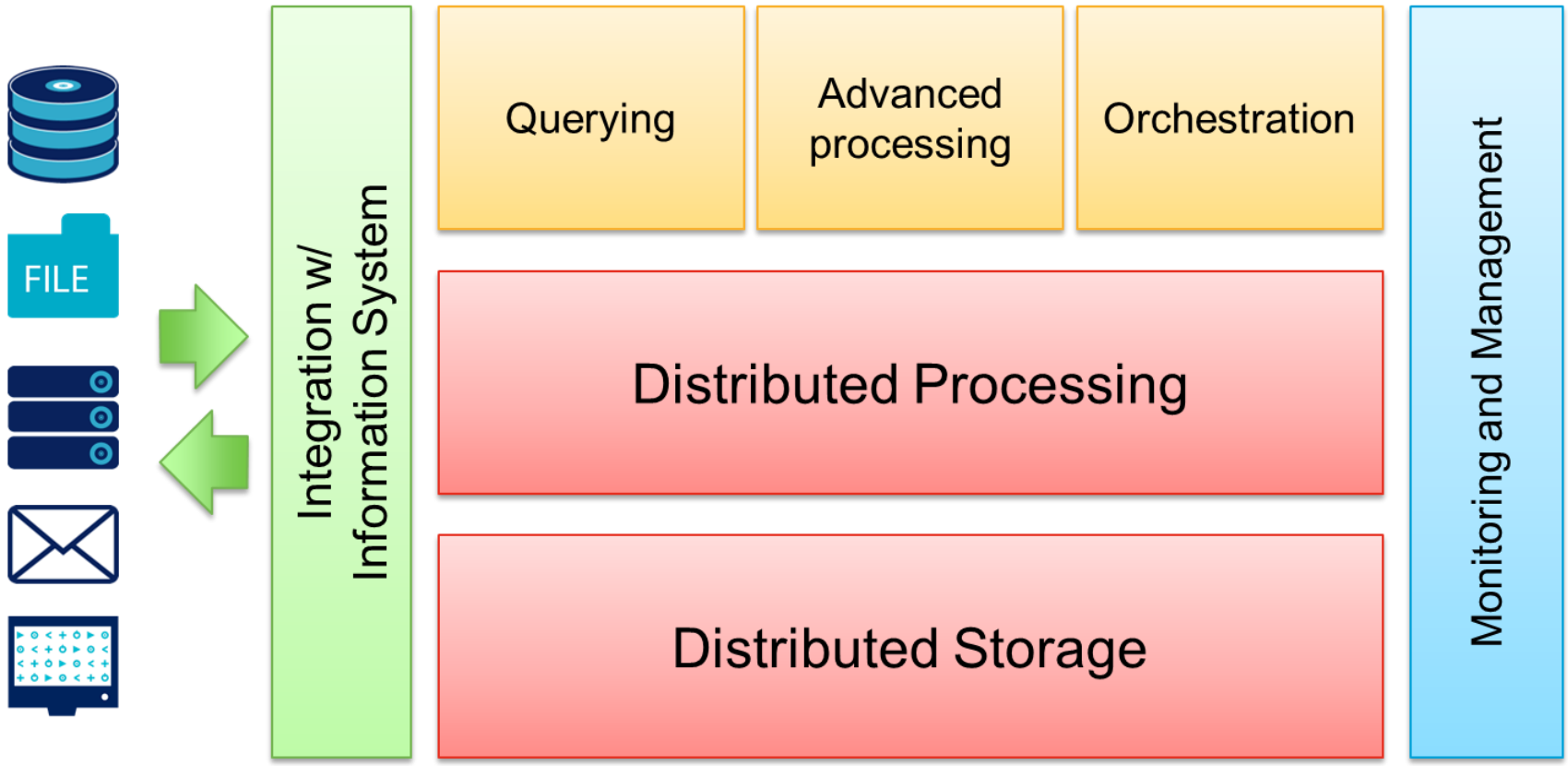
Enters Hadoop the Big Data Refinery

- Hadoop is not replacing anything.
- Hadoop has become another component in organizations enterprise data platform.
- Hadoop (Big Data Refinery) can ingest data from all types of different sources.
- Hadoop then interacts with traditional systems that provide transactions and interactions (relational databases) and business intelligence and analytic systems (data warehouses).



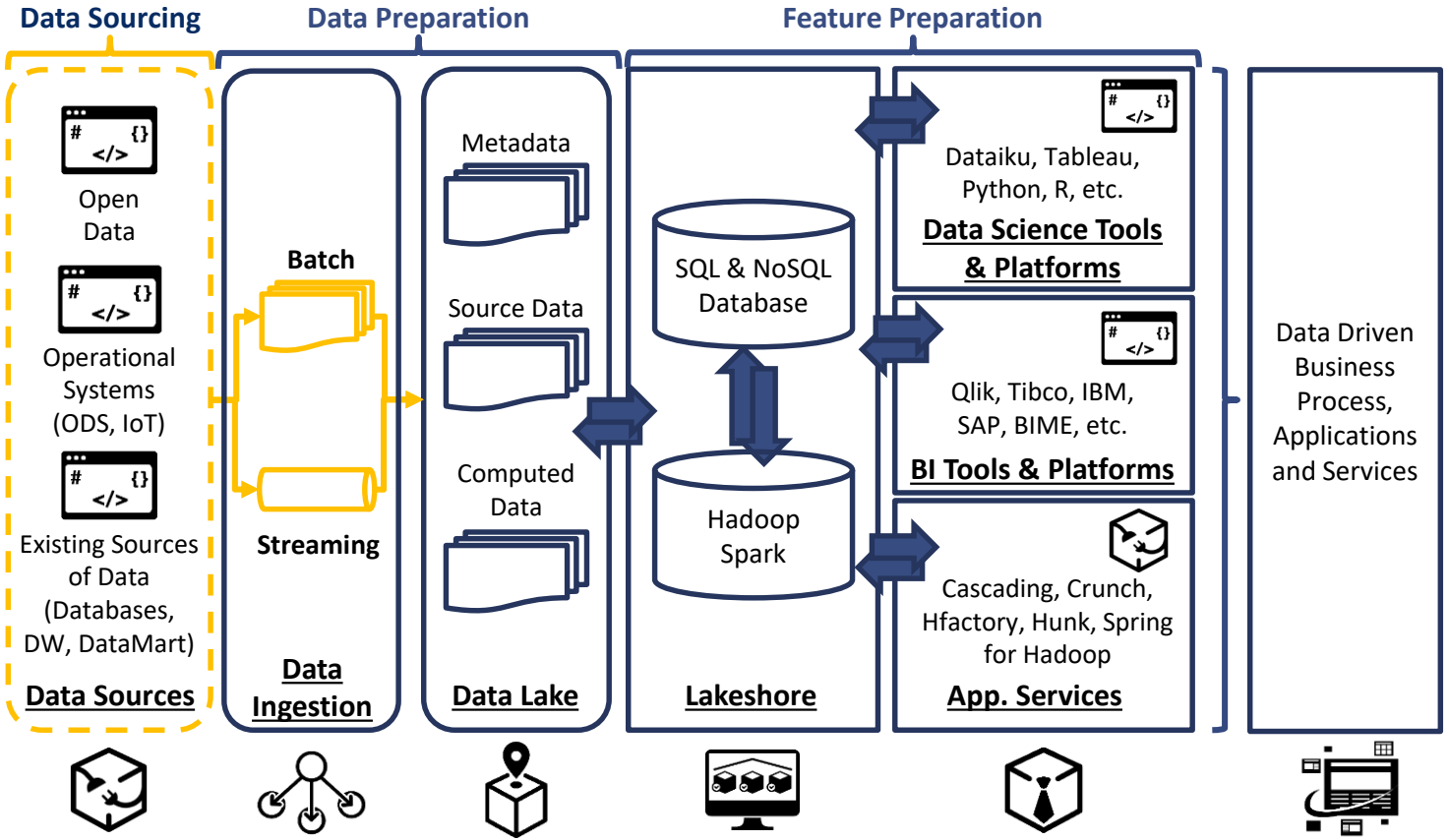
Source: DBA Journey Blog

Hadoop Platform

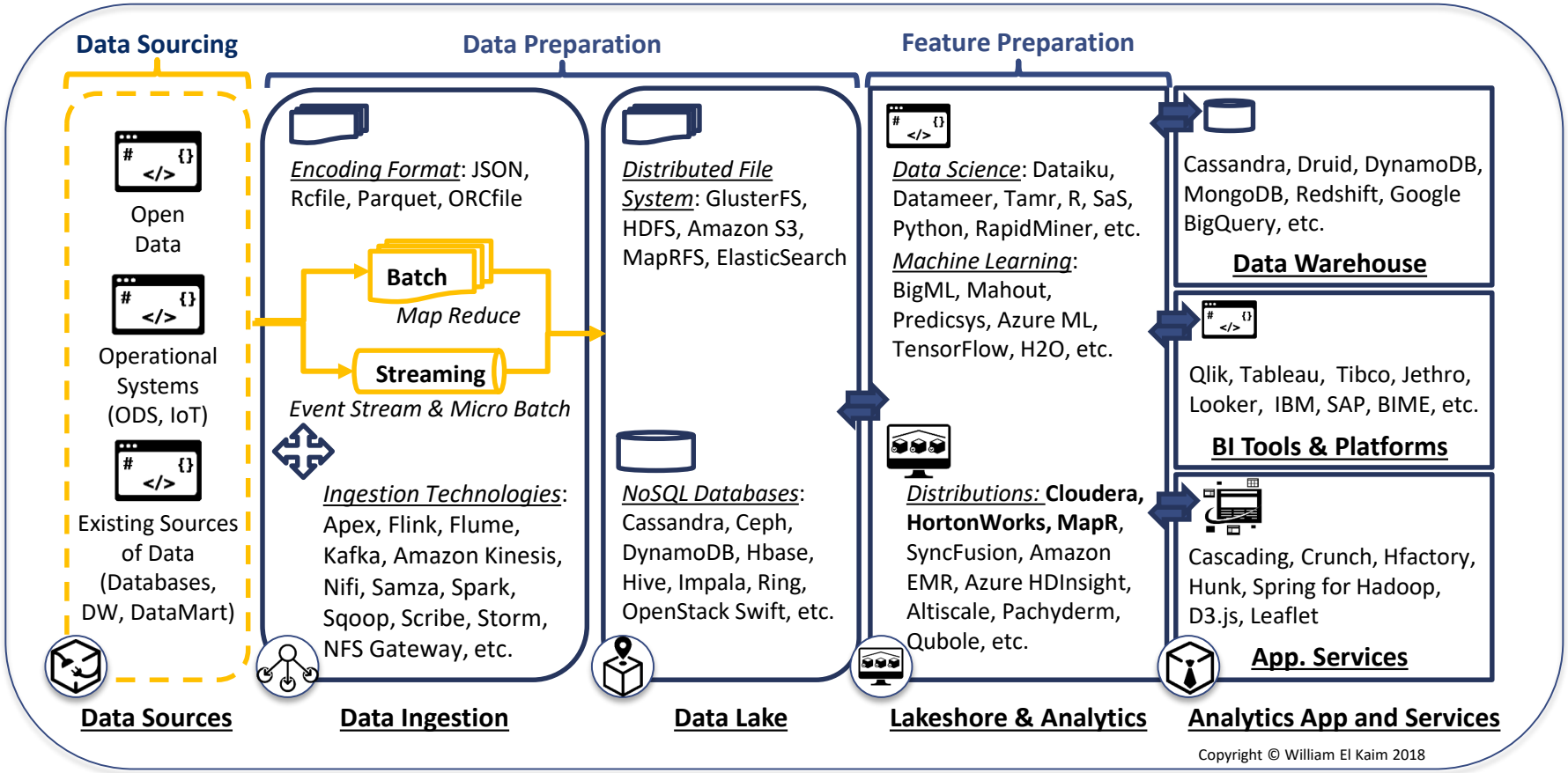


Source: Octo Technology

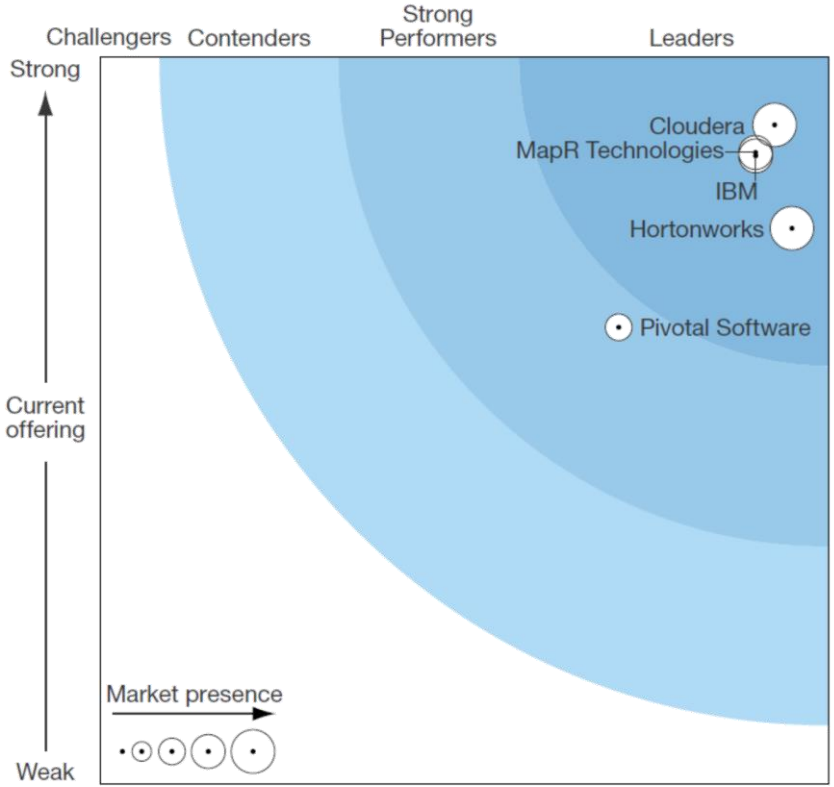
Hadoop Architecture



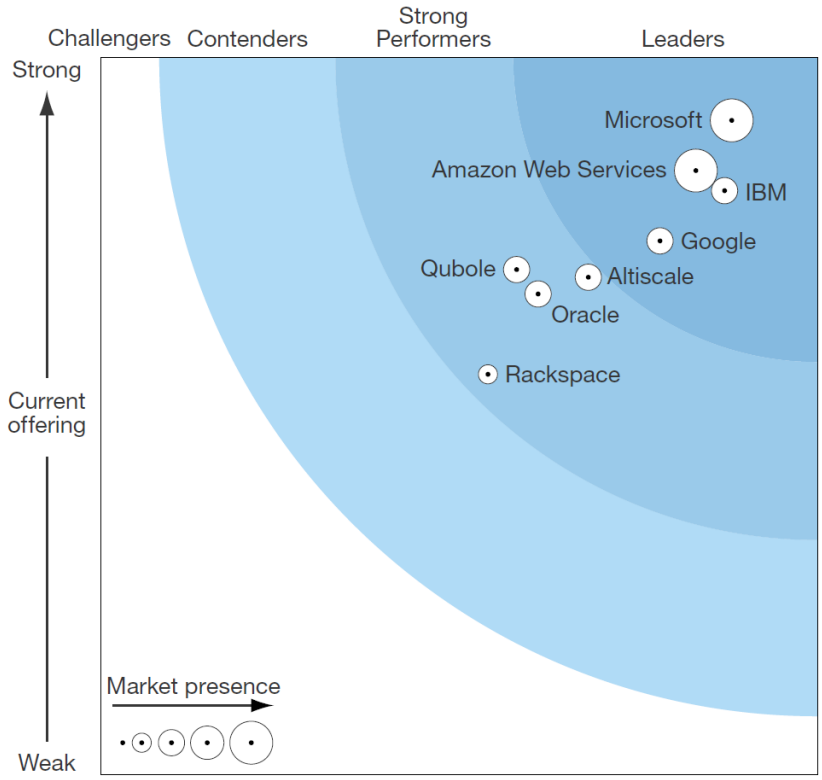
Hadoop Technologies



Hadoop Distributions and Providers



Forrester Big Data Hadoop Distributions, Q1 2016



Forrester Big Data Hadoop Cloud, Q1 2016

Hadoop Distributions and Providers

- Three Main pure-play Hadoop distributors
  - Cloudera, Hortonworks, and MapR Technologies
- Other Hadoop distributions
  - BigTop, Pachyderm, SyncFusion, Big Data Europe platform

	Hortonworks	Cloudera	MapR	Bigtop	BDE
File System	HDFS	HDFS	NFS	HDFS	HDFS
Installation	Native	Native	Native	Native	lightweight virtualization
Plug & play components (no rigid schema)	no	no	no	no	yes
High Availability	Single failure recovery (yarn)	Single failure recovery (yarn)	Self healing,mult. failure rec.	Single failure recovery (yarn)	Multiple Failure recovery
Cost	Commercial	Commercial	Commercial	Free	Free
Scaling	Freemium	Freemium	Freemium	Free	Free
Addition of custom components	Not easy	No	No	No	Yes
Integration testing	yes	yes	yes	yes	--
Operating systems	Linux	Linux	Linux	Linux	All
Management tool	Ambari	Cloudera manager	MapR Control system	-	Dockerswarm UI+ Custom

Source: Big Data Europe

Hadoop Distributions and Providers

- Hadoop Cloud Providers
  - Amazon EMR, BigStep, BlueData, Datastax Enterprise Analytics, Google Cloud DataProc, IBM BigInsights, Microsoft HDInsight, Oracle Big Data, Packet, Qubole, Rackspace, SAP Cloud Platform Big Data Services, Teradata Enterprise Access for Hadoop
- Big Data Factories
  - Dataiku, Saagie,

Enterprise Datawarehouse Limitations

- Organizations realized that traditional Enterprise Datawarehouse (EDW) technologies can't meet their new business needs
  - Including leveraging streaming and social data from the Web or from connected devices on the Internet of things (IoT)
- One particular challenge with traditional Enterprise Datawarehouse was their schema-on-write architecture
  - In an EDW you must design the data model and articulate the analytic frameworks before loading any data.
  - In other words, you need to know ahead of time how you plan to use that data.

Source: Zaloni

Schema On Write …

- Concept: Before any data is written in the database, the structure of that data is strictly defined, and that metadata stored and tracked.
- Irrelevant data is discarded, data types, lengths and positions are all delineated.
  - The schema; the columns, rows, tables and relationships are all defined first for the specific purpose that database will serve.
  - Then the data is filled into its pre-defined positions. The data must all be cleansed, transformed and made to fit in that structure before it can be stored in a process generally referred to as ETL (Extract Transform Load).
- That is why it is called “schema on write” because the data structure is already defined when the data is written and stored.
- For a very long time, it was believed that this was the only right way to manage data.

Source: AdaptiveSystems

Schema On Write …

- Benefits: Quality and Query Speed.
  - Because the data structure is defined ahead of time, when you query, you know exactly where your data is.
  - The structure is generally optimized for the fastest possible return of data for the types of questions the data store was designed to answer (write very simple SQL and get back very fast answers).
  - The answers received from querying data are sharply defined, precise and trustworthy, with little margin for error.
- Drawbacks: Data Alteration & Query Limitations
  - Data has been altered and structured specifically to serve a specific purpose. Chances are high that, if another purpose is found for that data, the data store will not suit it well.
  - ETL processes and validation rules are then needed to clean, de-dupe, check and transform that data to match pre-defined format. Those processes take time to build, time to execute, and time to alter if you need to change it to suit a different purpose.

Source: AdaptiveSystems

Schema On Read … and Hadoop

- Revolutionary concept: "You don't have to know what you're going to do with your data before you store it."
  - Data of many types, sizes, shapes and structures can all be thrown into the Hadoop Distributed File System, and other Hadoop data storage systems.
  - While some metadata needs to be stored, to know what's in there, no need yet to know how it will be structured!
- Therefore, the data is stored in its original granular form, with nothing thrown away
  - In fact, no structural information is defined at all when the data is stored.
- So "schema on read" implies that the schema is defined at the time the data is read and used, not at the time that it is written and stored.
  - When someone is ready to use that data, then, at that time, they define what pieces are essential to their purpose, where to find those pieces of information that matter for that purpose, and which pieces of the data set to ignore.

Source: AdaptiveSystems

Schema On Read … and Hadoop

- Benefits: Flexibility and Query Power
  - Because data is stored in its original form, nothing is discarded, or altered for a specific purpose.
  - Different types of data generated by different sources can be stored in the same place. This allows you to query multiple data stores and types at once.
    - The heart of the Hadoop data lake concept which puts all available data sets in their original form in a single location such a potent one.
- Drawbacks: Inaccuracies and Slow Query Speed
  - Since the data is not subjected to rigorous ETL and data cleansing processes, nor does it pass through any validation, data may be riddled with missing or invalid data, duplicates and a bunch of other problems that may lead to inaccurate or incomplete query results.
  - In addition, since the structure must be defined when the data is queried, the SQL queries tend to be very complex. They take time to write, and even more time to execute.

Source: AdaptiveSystems

Schema On Read or Schema On Write?

- Schema on read options tend to be a better choice:
  - for exploration, for “unknown unknowns,” when you don’t know what kind of questions you might want to ask, or the kinds of questions might change over time.
  - when you don’t have a strong need for immediate responses. They’re ideal for data exploration projects, and looking for new insights with no specific goal in mind.
- Schema on write options tend to be very efficient for “known unknowns.”
  - When you know what questions you’re going to need to ask, especially if you will need the answers fast, schema on write is the only sensible way to go.
  - This strategy works best for old school BI types of scenarios on new school big data sets.

Source: AdaptiveSystems

The Hadoop Data Lake Concept

- The Hadoop data lake concept can be summed up as, "Store it all in one place, figure out what to do with it later"
- While this might be the general idea of your Hadoop data lake, you won't get any real value out of that data until you figure out a logical structure for it.
- And you'd better keep track of your metadata one way or another.
  - Risk: a lake full of data, but no idea what lies under the shiny surface.
- At some point, you have to give that data a schema, especially if you want to query it with SQL or something like it.
- The eternal Hadoop question is whether to apply the brave new strategy of schema on read, or to stick with the tried and true method of schema on write.

Source: AdaptiveSystems

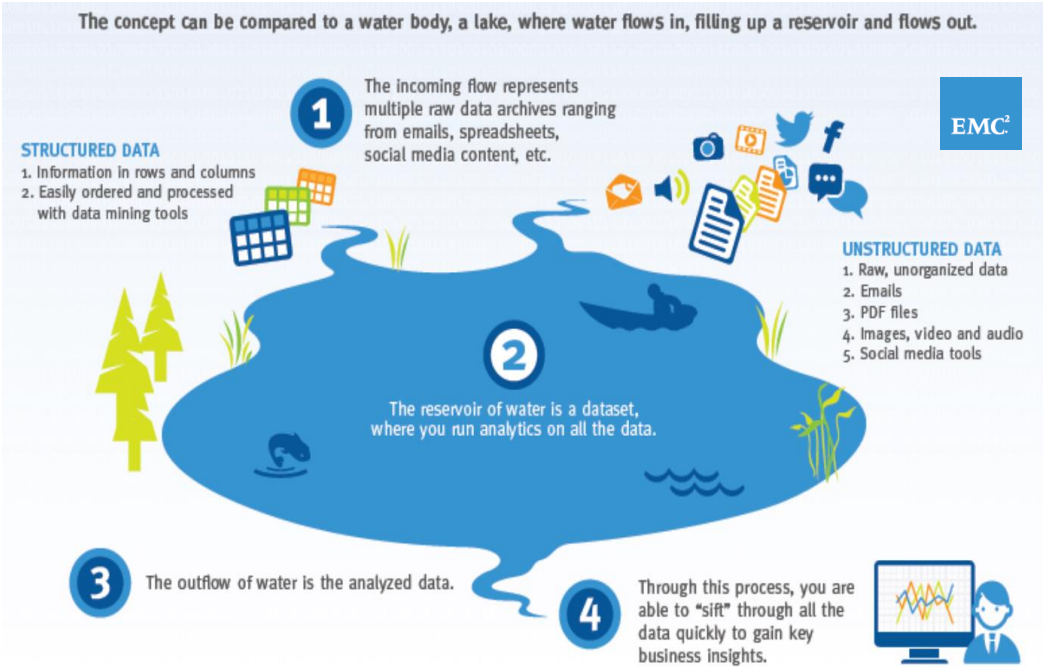
Hadoop to the Rescue to offer

- Unified distributed fault-tolerant and scalable storage
- Cost Effective.
  - Hadoop can be 10 to 100 times less expensive to deploy than traditional data warehouse.
- Just-in-time ad'hoc dataset creation.
  - Extract and place data into an Hadoop-based repository without first transforming the data (as you would do for a traditional EDW).
  - All jobs are created in an ad hoc manner, with little or no required preparation work.
  - Labor intensive processes of cleaning up data and developing schema are deferred until a clear business need is identified.
- Data governance for structured and unstructured "raw" data

Source: Zaloni

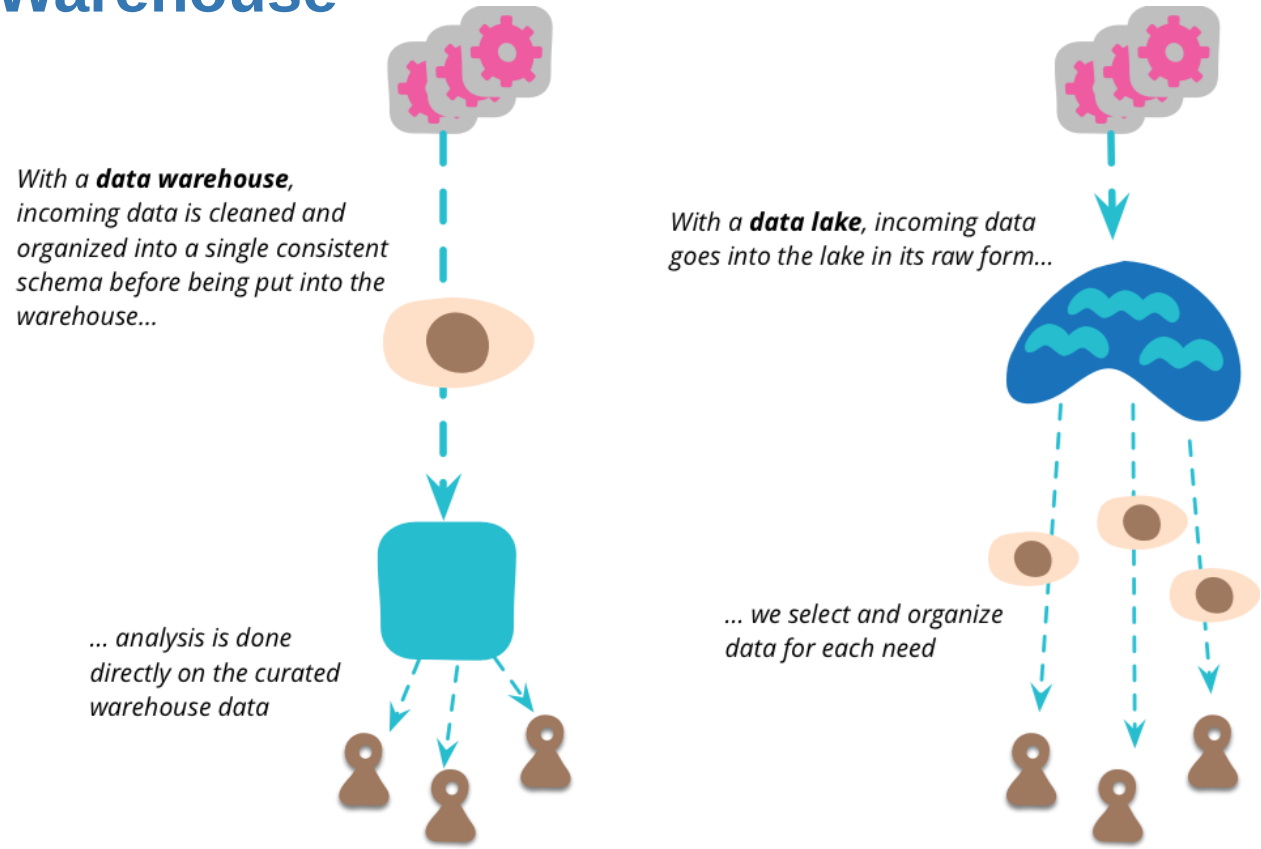
Hadoop Data Lake

- Supports structured or unstructured data.
- Benefiting from a variety of storage and processing tools to extract value quickly.
- Requiring little or no processing for adapting the structure to an enterprise schema
- A central location in which to store all your data in its native form, regardless of its source or format.



Source: Zaloni

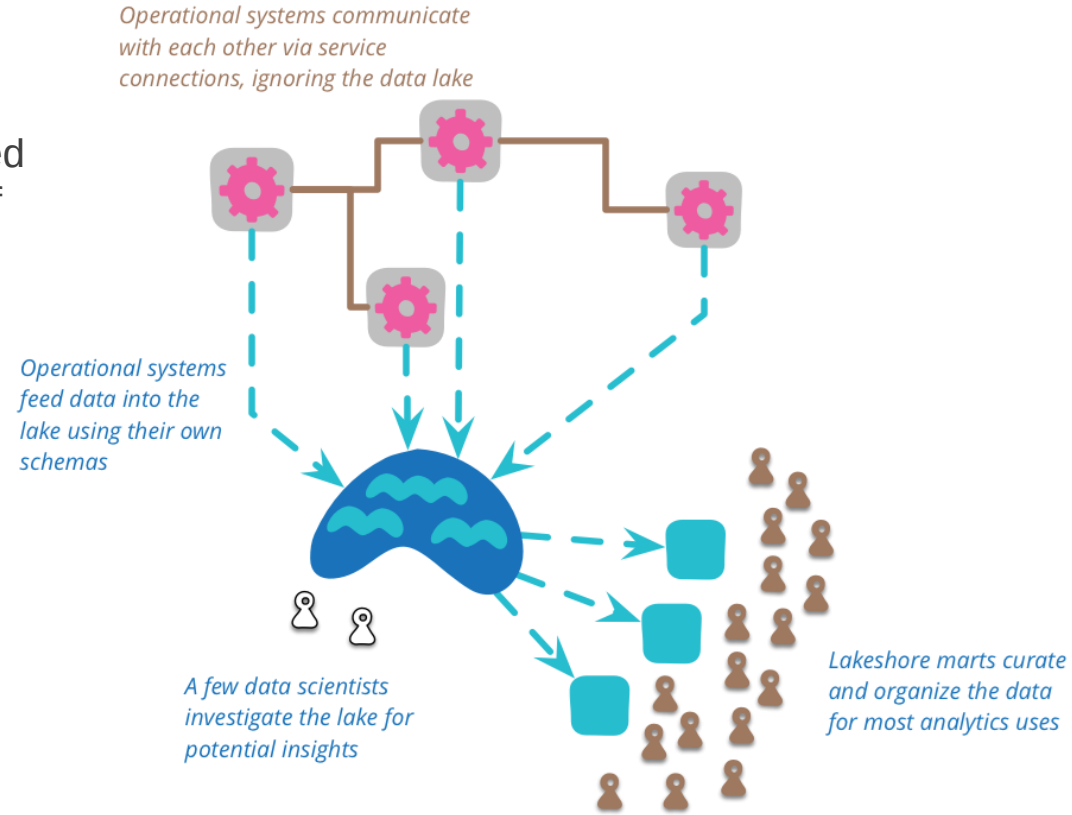
Hadoop Data Lake vs. Enterprise Data Warehouse



Source: Martin Fowler

The Lakeshore Concept

- The data lake shouldn't be accessed directly very much.
  - Because the data is raw, you need a lot of skill to make any sense of it.
- Lakeshore
  - Create a number of data marts each of which has a specific model for a single bounded context.
  - A larger number of downstream users can then treat these lakeshore marts as an authoritative source for that context.



Source: Martin Fowler

Data Lake Tools

- Bigstep
- Denodo (Virtual Data Lake using data virtualization)
- Informatica Data Lake Mgt & Intelligent Data Lake
- Microsoft Azure Data Lake and Azure Data Catalog
- Koverse
- Oracle
- Platfora
- Podium Data
- Waterline Data Inventory
- Zaloni Bedrock
- Zdata Data Lake

Data Lake Resources

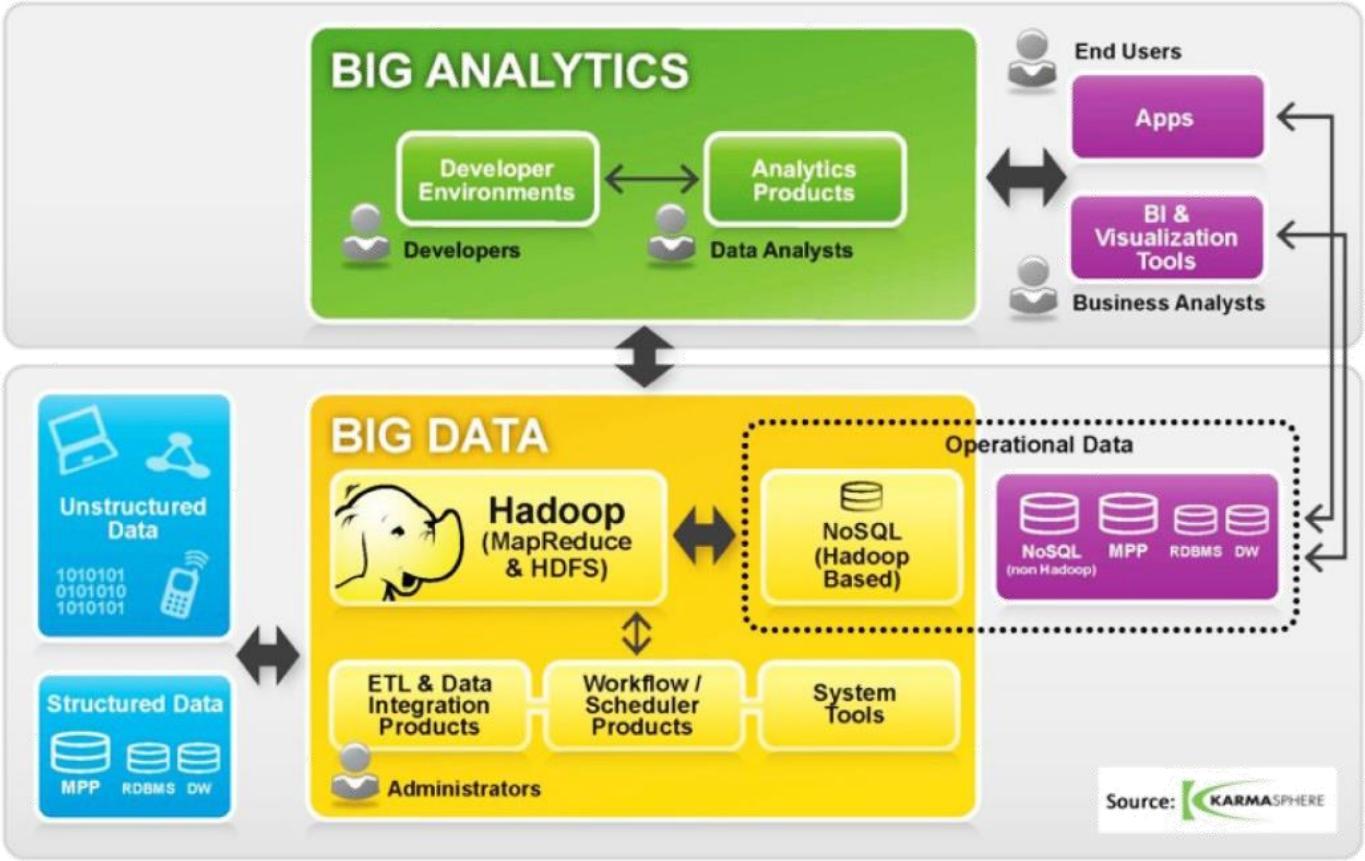
- PWC: Data lakes and the promise of unsiloed data
- Zaloni: Resources

When to Use Hadoop?

Two Main Use Cases

2 **Big Data Analytics and Use**

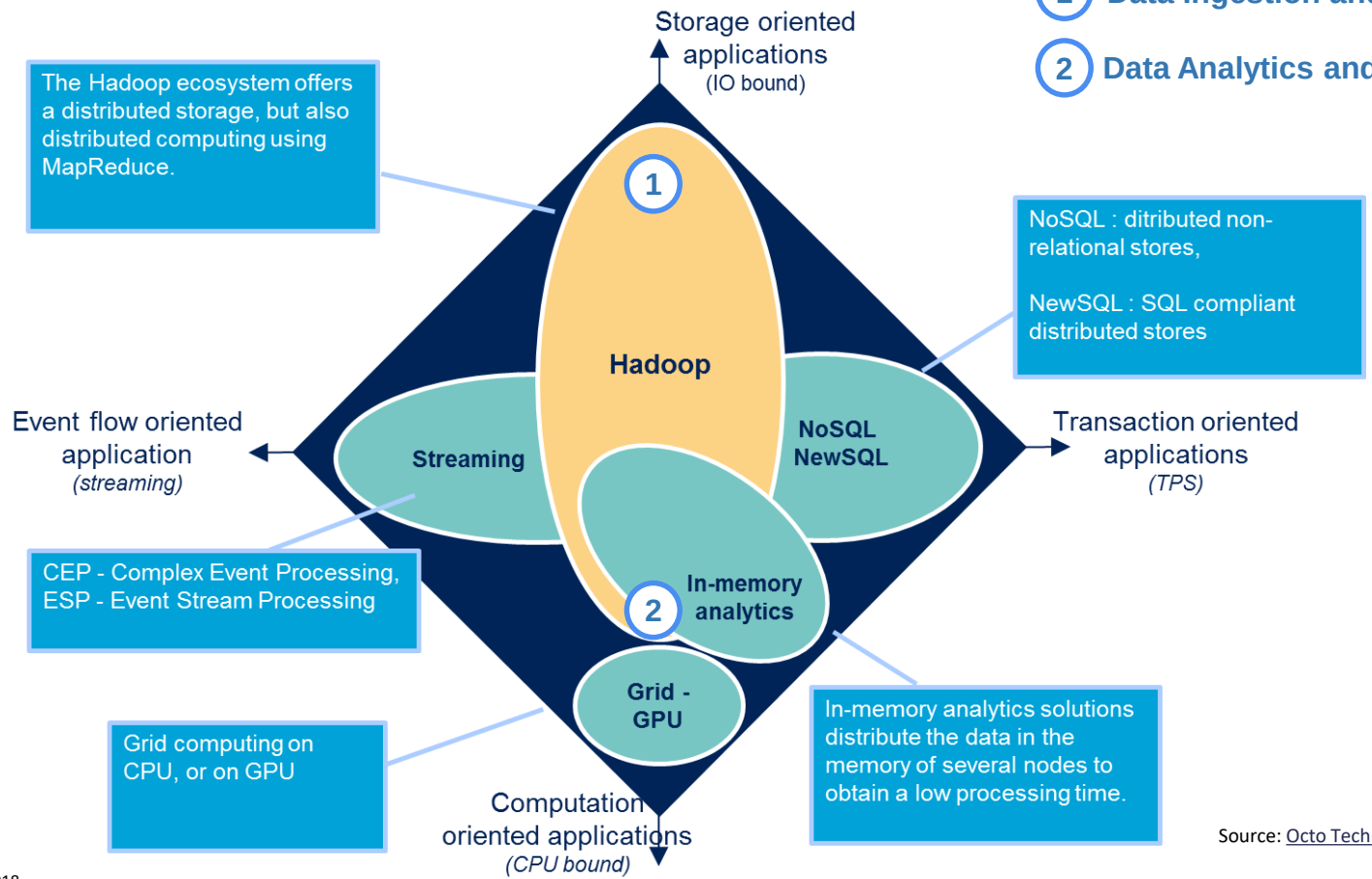
1 **Data Ingestion and Storage**



When to Use Hadoop?

1 Data Ingestion and Storage

2 Data Analytics and Use

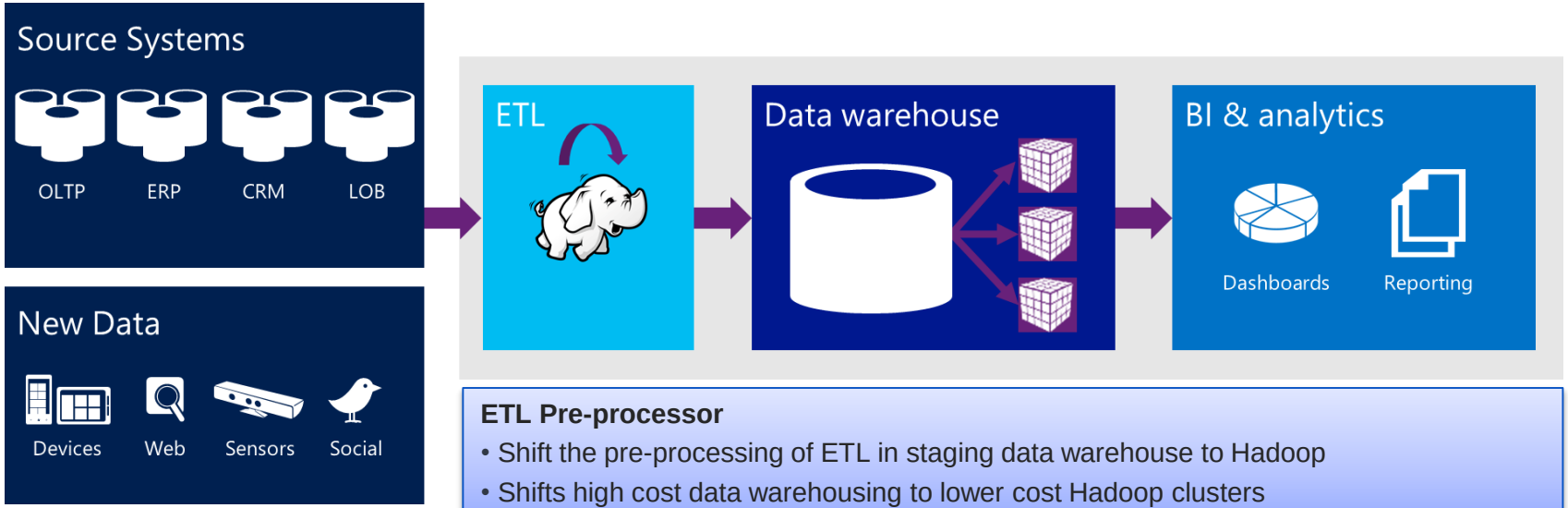


Source: Octo Technology

Hadoop for Data Mgt and Storage

ETL Pre-processor

1 Data Ingestion and Storage



ETL Pre-processor

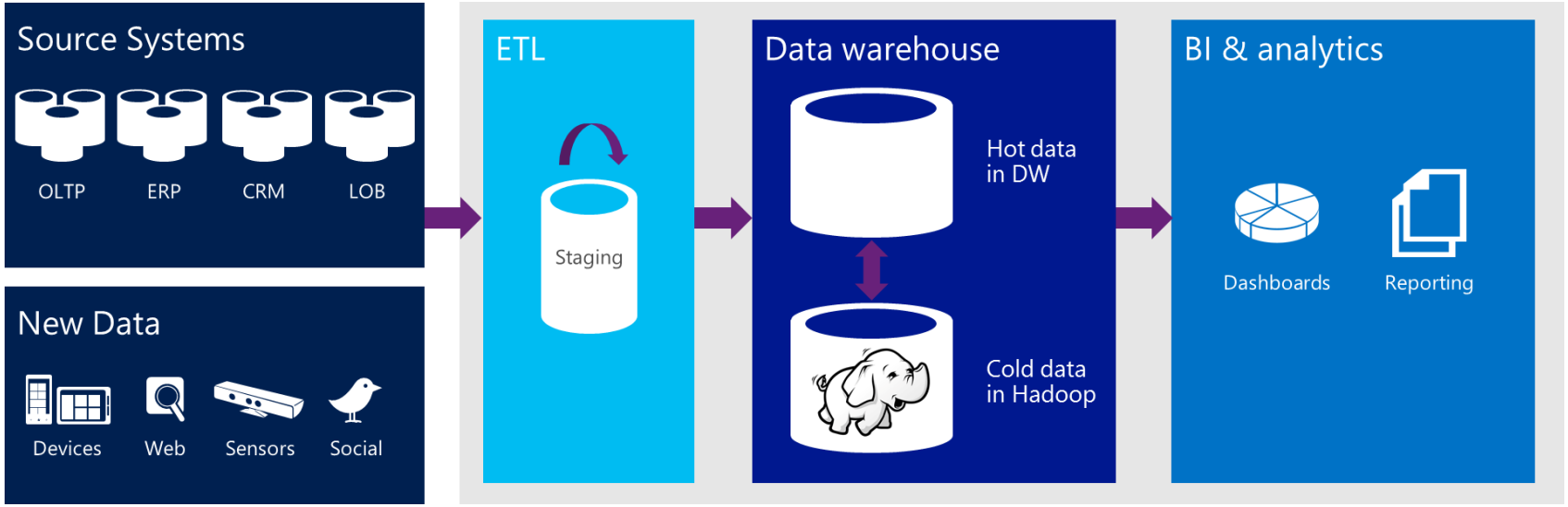
- Shift the pre-processing of ETL in staging data warehouse to Hadoop
- Shifts high cost data warehousing to lower cost Hadoop clusters

Source: Microsoft

Hadoop for Data Mgt and Storage

Massive Storage

1 Data Ingestion and Storage



Massive Storage

- Offloading large volume of historical data into cold storage with Hadoop
- Keep data warehouse for hot data to allow BI and analytics
- When data from cold storage is needed, it can be moved back into the warehouse

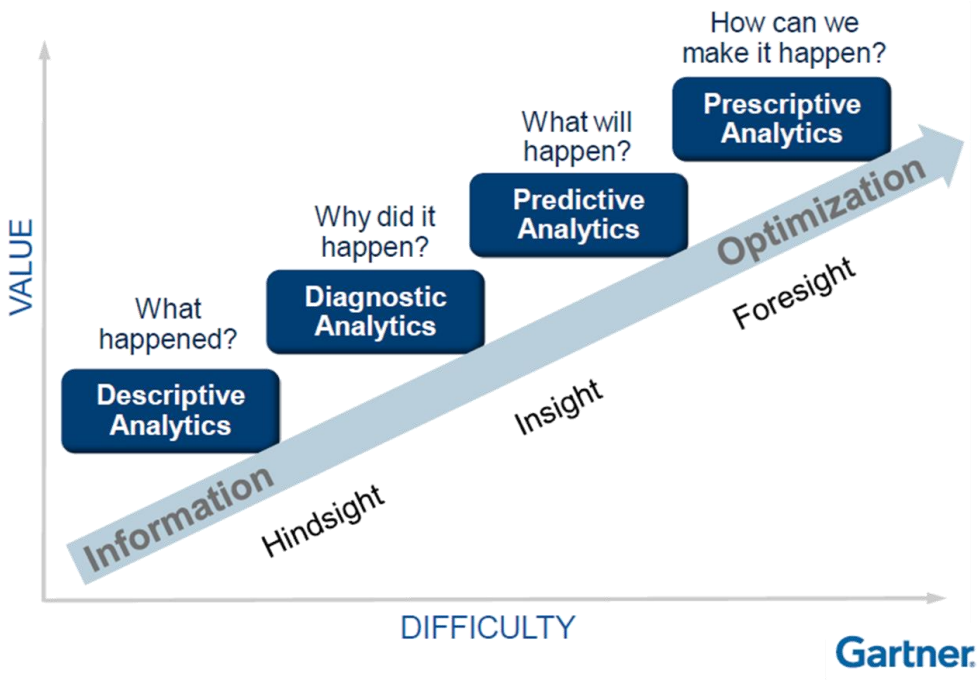
Source: Microsoft

Hadoop for Data Analytics and Use

Six V to Nirvana to Provide

2 Data Analytics and Use

- Hindsight (what happened?)
- Oversight (what is happening?)
- Insight (why is it happening?)
- Foresight (what will happen?)



Hadoop for Data Analytics and Use

From Hindsight to Insight to Foresight

2 Data Analytics and Use

- Traditional analytics tools are not well suited to capturing the full value of big data.
  - The volume of data is too large for comprehensive analysis.
  - The range of potential correlations and relationships between disparate data sources are too great for any analyst to test all hypotheses and derive all the value buried in the data.
  - Basic analytical methods used in business intelligence and enterprise reporting tools reduce to reporting sums, counts, simple averages and running SQL queries.
  - Online analytical processing is merely a systematized extension of these basic analytics that still rely on a human to direct activities specify what should be calculated.

Hadoop for Data Analytics and Use

From Hindsight to Insight to Foresight

2 Data Analytics and Use

From traditional reporting…	… to Business Intelligence…	… to Analytics and Big Data
"Push"	"Pull"	"Predictive"
Mostly fixed format	More interactive / self service (drill down, slice-and-dice, etc)	Interactive and fully business led
Typically finance-centric offering little to other functions	Applies to all business functions, 'front office' (customer facing) and 'back office' (finance, HR,etc)	Applies mostly to front office – often customer relationships and product development
Internal & structured data only, with little or no external collaboration	Still mostly internal and structured data, but bringing together more data sources (crossing internal silos of function and geography)	Combines internal and significant external data (e.g. spatial, statistical, 3rd party data); often unstructured (e.g. social media), and often using very large data sets
Implemented as an after-thought to ERP programmes	Implemented independently as a peer of ERP programmes (not subservient)	Implemented as a business capability, with dedicated analytic team(s) built into the organisation
Technology generally not a differentiator	More technology differentiation and choice, generally still 'on premise' (not as-a-Service)	Many specialist, highly differentiating technologies and tools; increasing use of open-source and cloud-based approaches.
Backward looking / "rear view mirror" for *control* (what happened?)	Still backward looking, but looking at *causality* (what happened and why?)	Forward looking / "head up display", looking at *correlations* to predict future outcomes (what will happen?)
Focussed on controlling and sustaining the business (bottom line)	Some elements of bottom and top line focus	Focussed on competitive advantage and growing the business (top line)

Hadoop for Data Analytics and Use

From Data Management To Data Driven Decisions

2 Data Analytics and Use

- Machine Learning: Data Reliability For Big Data
  - In big data world bringing data together from multiple internal and external sources can be a challenge. New approach to moving from manual or rules-based matching to matching done via machine learning.
  - The initial phase of machine learning being then to provide transparency of the actual rules that drive the merge and then it is up to the user to evaluate the discovered rule and persist it in the system.
- Graph: Finding Relationships In The Data
  - Graph is used to helping understand and navigate many-to-many relationships across all data entities.
- Cognitive Systems: Intelligent Recommendations
  - Building intelligent systems that guide users and provide intelligent recommendations, based on data and user behavior.

Source: Reltio

Hadoop for Data Analytics and Use

From Data Management To Data Driven Decisions

2 Data Analytics and Use

- Collaborative Curation: Clean & Current Data
  - Sharing data across all systems and functional groups helps realize the full value of data collected. Marketing, sales, services, and support should all leverage the same reliable, consolidated data. They should be able to collaborate and contribute to enriching the data. They should be able to vote on data quality or the business impact of any data entity. New data-driven applications must support this.
- Data Monetization & DaaS: New Revenue Streams
  - Charter of a CDO is not only about data governance and data integration and management. Increasingly, companies are asking CDOs to turn this data into new revenue streams. With cloud-based data-as-a-service, companies can monetize their data and become data brokers. Businesses can now collaborate with each other to create common data resources and easily share or exchange data.

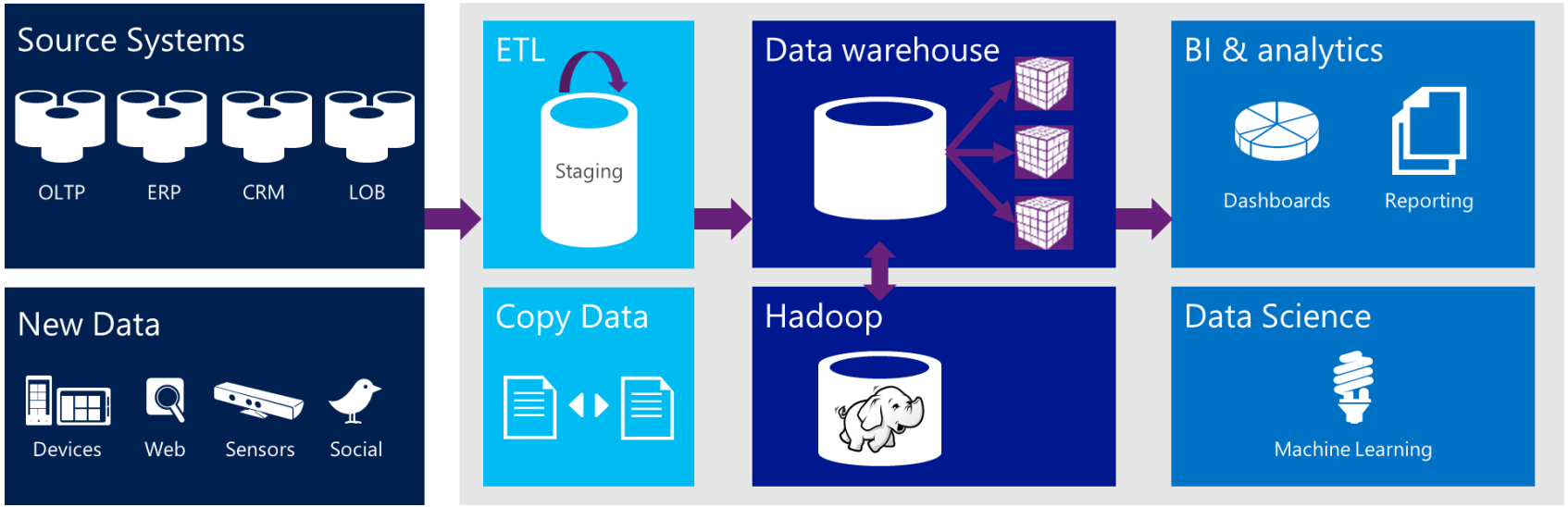
Source: Reltio

Hadoop for Data Analytics and Use

Data Discovery

2 Data Analytics and Use

2 Data Analytics and Use



Data Discovery

- Keep data warehouse for operational BI and analytics
- Allow data scientists to gain new discoveries on raw data (no format or structure)
- Operationalize discoveries back into the warehouse

Source: Microsoft

Hadoop for Data Analytics and Use

Skills Needed

2 Data Analytics and Use

ROLES	RESPONSIBILITES
Data Architect	Develops data architecture to effectively capture, integrate, organize, centralize and maintain data. Core responsibilities include: ✔ Data Warehousing Solutions ✔ Extraction, Transformation and Load (ETL) ✔ Data Architecture Development ✔ Data Modeling
Data Engineer	Develop, test and maintain data architectures to keep data accessible and ready for analysis. Key tasks are: ✔ Extraction Transformation and Load (ETL) ✔ Installing Data Warehousing Solutions ✔ Data Modeling ✔ Data Architecture Construction and Development ✔ Database Architecture Testing

Sources:
KDnuggets – www.kdnuggets.com/2015/11/different-data-science-roles-industry.html
Udacity – blog.udacity.com/2014/12/data-analyst-vs-data-scientist-vs-data-engineer.html
RJMetrics – rjmetrics.com/resources/reports/the-state-of-data-science/

BetterBuys

ROLES	RESPONSIBILITES
Data Analyst	Processes and interprets data to get actionable insights for a company. Responsibilities include: ✔ Data Collection and Processing ✔ Programming ✔ Machine Learning ✔ Data Munging ✔ Data Visualization ✔ Applying Statistical Analysis
Data Scientist	Data analysis once data volume and velocity reaches a level requiring sophisticated technical skills. Core tasks are: ✔ Data Cleansing and Processing ✔ Predictive Modeling ✔ Machine Learning ✔ Identifying Questions ✔ Running Queries ✔ Applying Statistical Analysis ✔ Correlating Disparate Data ✔ Storytelling and Visualization

39

When Not to Use Hadoop

- Real Time Analytics (Solved in Hadoop V2)
  - Since Hadoop V1 cannot be used for real time analytics, people explored and developed a new way in which they can use the strength of Hadoop (HDFS) and make the processing real time. So, the industry accepted way is to store the Big Data in HDFS and mount Spark over it. By using spark the processing can be done in real time and in a flash (real quick). Apache Kudu is also a complementary solution to use.
- To Replace Existing Infrastructure
  - All the historical big data can be stored in Hadoop HDFS and it can be processed and transformed into a structured manageable data. After processing the data in Hadoop you often need to send the output to other database technologies for BI, decision support, reporting etc.
- Small Datasets
  - Hadoop framework is not recommended for small-structured datasets as you have other tools available in market which can do this work quite easily and at a fast pace than Hadoop. For a small data analytics, Hadoop can be costlier than other tools.

Source: Edureka

Big Data Technologies

Ingestion

Ingestion Architecture:

- Scalable, Extensible to capture streaming and batch data.
- Provide capability to business logic, filters, validation, data quality, routing, etc. business requirements.

Technology Stack:

- Apache Flume
- Apache Kafka
- Apache Storm
- Apache Sqoop
- NFS Gateway

Storage/Retention

Data Storage:

- Depending on the requirements data is placed into Hadoop HDFS, Hive, Hbase, Elastic Search or In-memory.
- Metadata management
- Policy-based Data Retention is provided.

Technology Stack:

- HDFS
- Hive Tables
- Hbase/MapR DB
- Elastic Search

Processing

Data Processing:

- Processing is provided for both batch and near-realtime use cases
- Provision Workflows for repeatable Data processing
- Provide Late Data Arrival Handling

Technology Stack:

- Map Reduce
- Hive
- Spark
- Storm
- Drill

Access

Visualization and APIs:

- Dashboard and applications that provides valuable business insights
- Data will be made available to consumers using API, MQ Feed and DB access

Technology Stack:

- Qlik/Tableau/Spotfire
- REST APIs
- Apache Kafka
- JDBC

Management, Monitoring, Governance

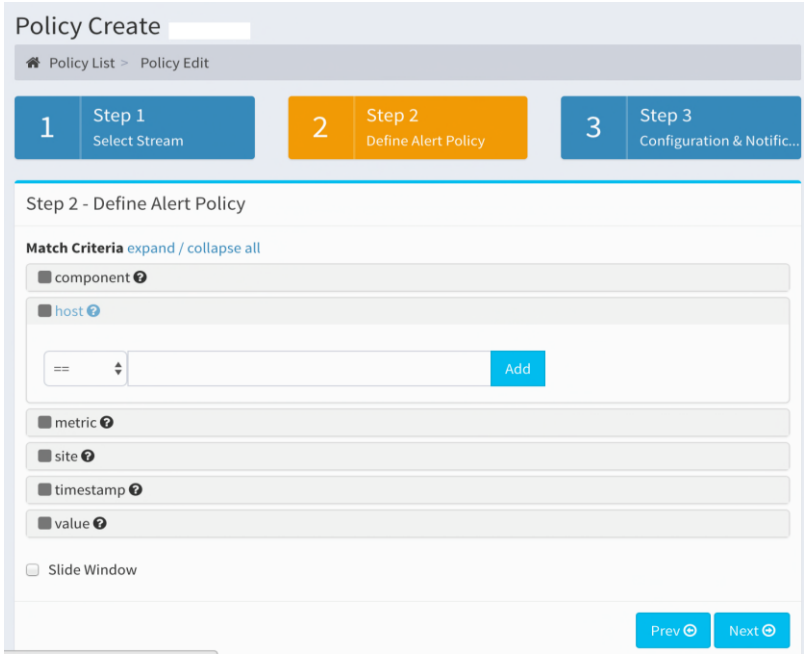
Ambari, Cloudera Manager, Cloudera Navigator, MapR MCS

Hadoop Security

- Apache Ranger is a framework to enable, monitor and manage comprehensive data security across the Hadoop platform.
  - With the advent of Apache YARN, the Hadoop platform can now support a true data lake architecture. Enterprises can potentially run multiple workloads, in a multi tenant environment.
- Apache Metron provides a scalable advanced security analytics framework built with the Hadoop Community evolving from the Cisco OpenSOC Project.
  - A cyber security application framework that provides organizations the ability to detect cyber anomalies and enable organizations to rapidly respond to identified anomalies.
- Apache Sentry is a system to enforce fine grained role based authorization to data and metadata stored on a Hadoop cluster.

Hadoop Security

- Apache Eagle: Analyze Big Data Platforms For Security and Performance
  - Apache Eagle is an Open Source Monitoring Platform for Hadoop ecosystem, which started with monitoring data activities in Hadoop.
  - It can instantly identify access to sensitive data, recognize attacks/malicious activities and blocks access in real time.
  - In conjunction with components (such as Ranger, Sentry, Knox, DgSecure and Splunk etc.), Eagle provides comprehensive solution to secure sensitive data stored in Hadoop.
  - As of 0.3.0, Eagle stores metadata and statistics into HBASE, and support Druid as metric store.



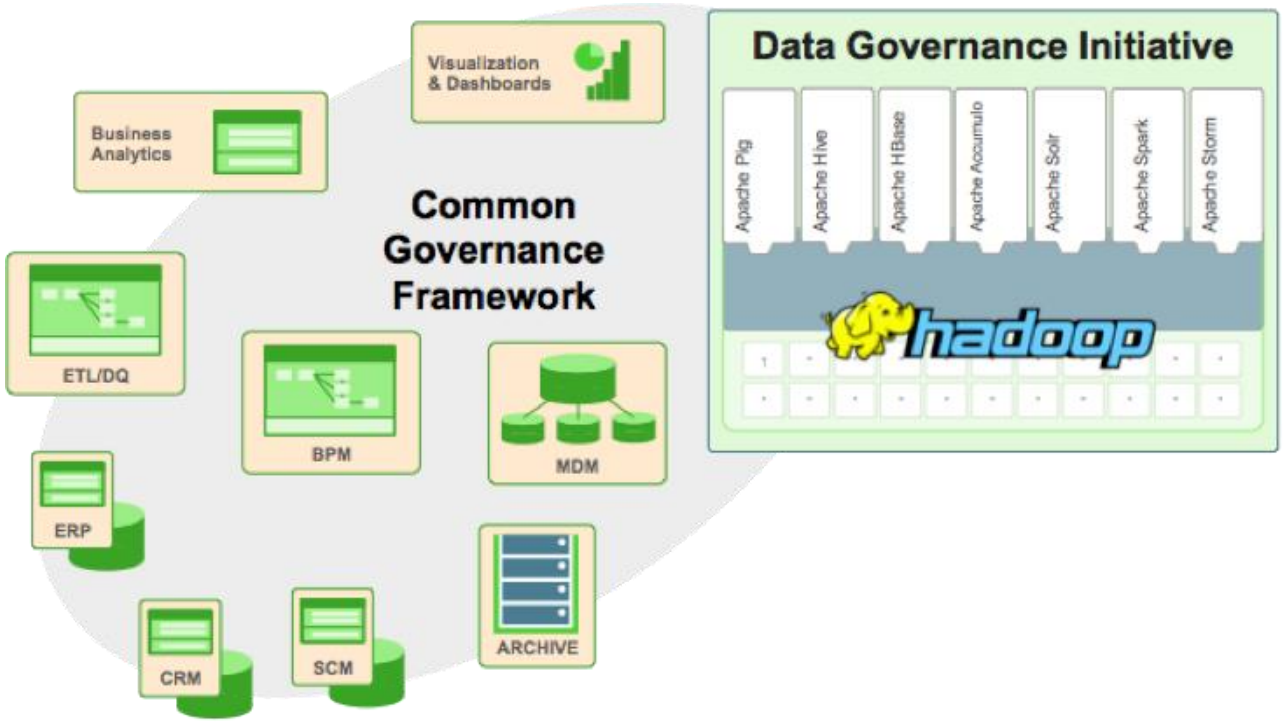
Hadoop Governance

Data Governance Initiative

- Enterprises adopting modern data architecture with Hadoop must reconcile data management realities when they bring existing and new data from disparate platforms under management.
- As customers deploy Hadoop into corporate data and processing environments, metadata and data governance must be vital parts of any enterprise-ready data lake.
- Data Governance Initiative (DGI)
  - with Aetna, Merck, Target, and SAS
  - Introduce a common approach to Hadoop data governance into the open source community.
  - Shared framework to shed light on how users access data within Hadoop while interoperating with and extending existing third-party data governance and management tools.
- A new project proposed to the apache software foundation: Apache Atlas

Hadoop Governance

Data Governance Initiative



TWO Requirements

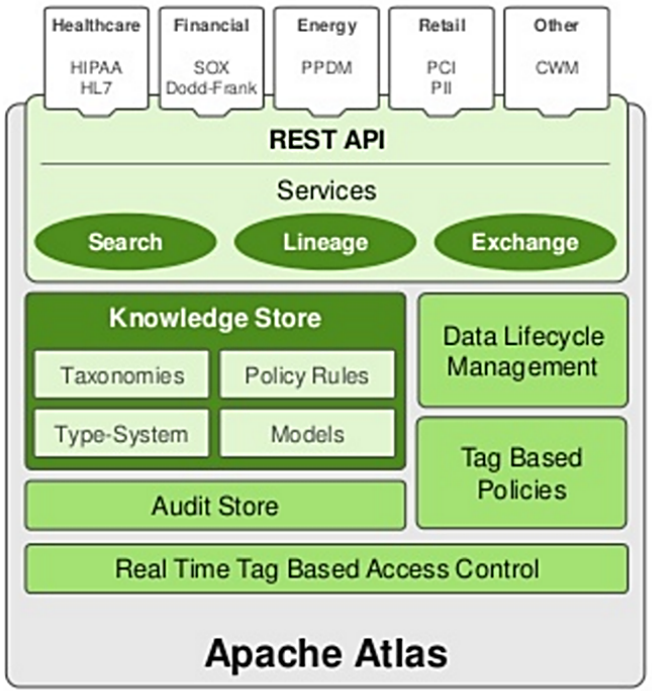
1. Hadoop must snap in to the existing frameworks and openly exchange metadata
2. Hadoop must address governance within its own stack of technologies

Hadoop Governance

Apache Atlas and Apache Falcon

- Apache Atlas is a scalable and extensible set of core foundational governance services
  - It enables enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the whole enterprise data ecosystem.
- Apache Falcon is a framework for managing data life cycle in Hadoop clusters
  - addresses enterprise challenges related to Hadoop data replication, business continuity, and lineage tracing by deploying a framework for data management and processing.
  - Falcon centrally manages the data lifecycle, facilitate quick data replication for business continuity and disaster recovery and provides a foundation for audit and compliance by tracking entity lineage and collection of audit logs.

Hadoop Governance
Apache Atlas Capabilities



Data Classification

- Import or define taxonomy business-oriented annotations for data
- Define, annotate, and automate capture of relationships between data sets and underlying elements including source, target, and derivation processes
- Export metadata to third-party systems

Centralized Auditing

- Capture security access information for every application, process, and interaction with data
- Capture the operational information for execution, steps, and activities

Search & Lineage (Browse)

- Pre-defined navigation paths to explore the data classification and audit information
- Text-based search features locates relevant data and audit event across Data Lake quickly and accurately
- Browse visualization of data set lineage allowing users to drill-down into operational, security, and provenance related information

Security & Policy Engine

- Rationalize compliance policy at runtime based on data classification schemes
- Advanced definition of policies for preventing data derivation based on classification (i.e. re-identification)

Source: Apache Atlas

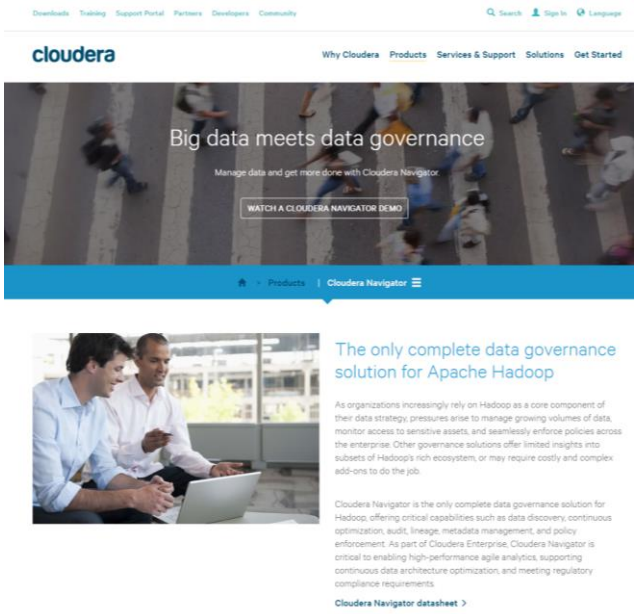
Hadoop Governance

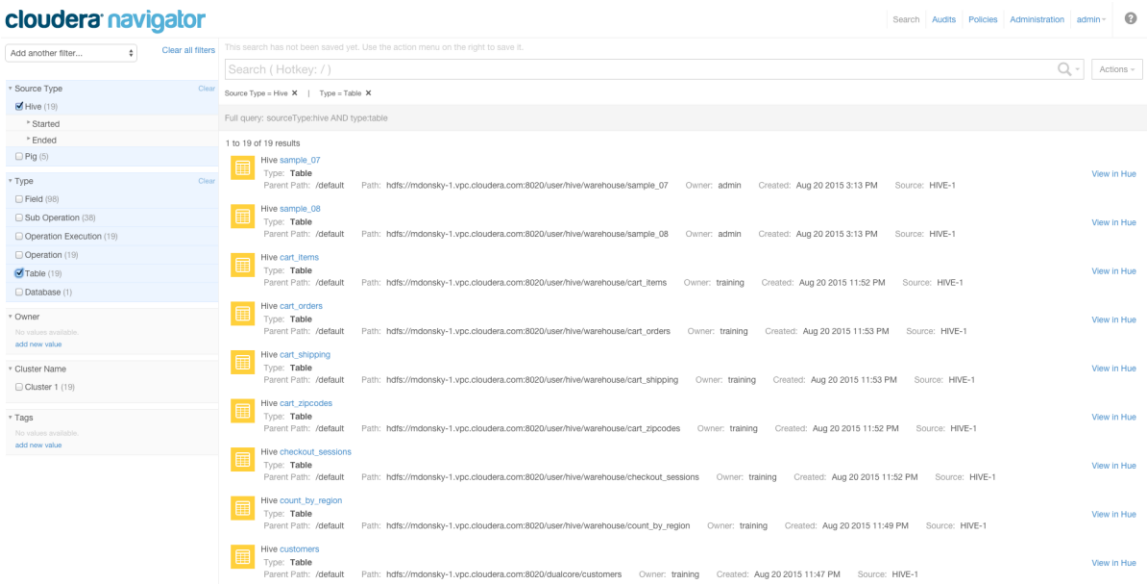
Other Vendors Entering The Market

- Alation
- Cloudera Navigator
- Collibra
- HortonWorks Data Platform
- Informatica Big Data Management
- mapR Converged Data Platform
- Podium Data
- Zaloni
- Zeenea

Hadoop Governance

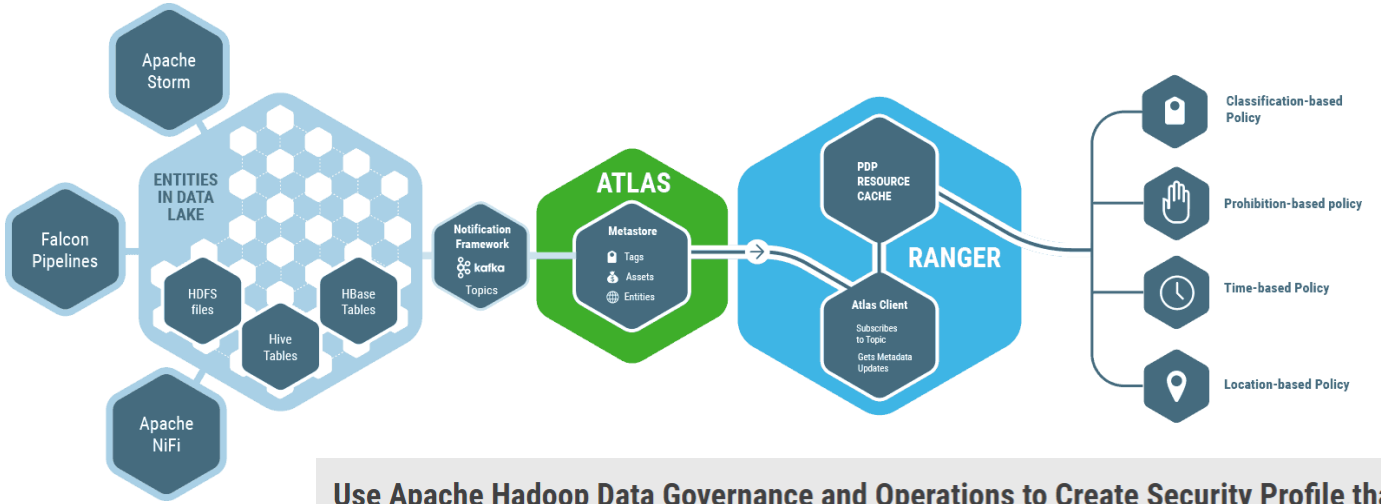
Cloudera Navigator





https://www.cloudera.com/products/cloudera-navigator.html

HortonWorks Data Platform Governance



Use Apache Hadoop Data Governance and Operations to Create Security Profile that Meet the Needs of Data-Driven Enterprises

Classification-based Policy
A data asset such as a table or column can be marked with the metadata tag related to compliance or business taxonomy (such as "PCI"). This tag is then used to assign permission to a user group.

Location-based Policy
Administrators can customize entitlements based on geography. A user trying to access the same data from different locations would be subject to unique geographical context thereby triggering access based on different set of privacy rules.

Data Expiry-based Policy
Apache Atlas can assign expiration dates to a data tag. Apache Ranger would inherit the expiration date and automatically deny users access to the tagged data after the expiration date. This policy is relevant for business use cases where data can become toxic after an expiry date.

Prohibition-based Policy
It is now possible to define a security policy that restricts combining two data sets. Administrators can now apply a metadata tag to both data sets to prevent them from being combined, helping avoid privacy violations.

Mapr MCS

With MapR Control System (MCS), you can manage any volume, table, or stream from a single interface. MCS gives you a single pane of glass for cluster metrics, alarms, and service logs as well as a curated user experience with streamlined workflows for common user actions.

MCS is built on top of a scalable and secure monitoring framework and can be deployed anywhere – on-premises, edge, or cloud.

New Dashboard with Event Co-Relation and Actionable Recommendation

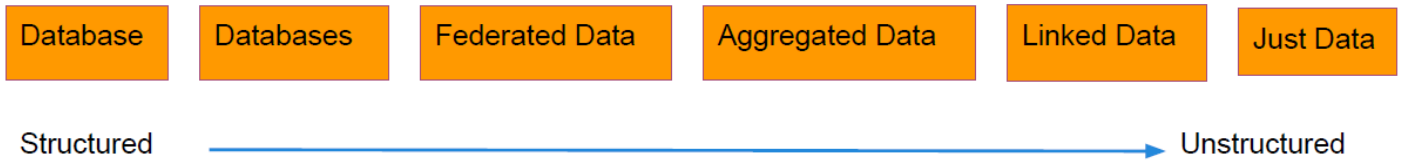
The cluster dashboard is often the most used part the application. In MCS, the dashboard offers a birds-eye view of critical cluster information that includes utilization and resource metrics, YARN statistics, node health by service and by topology, and active alarms.

The active alarms are now part of the metric chart timeline that helps administrators identify the data points around critical cluster events. The details on each alarm list the entities affected, description of the alarm, and also recommended next steps to resolve.

https://mapr.com/products/mapr-control-system/

Big Data Challenges

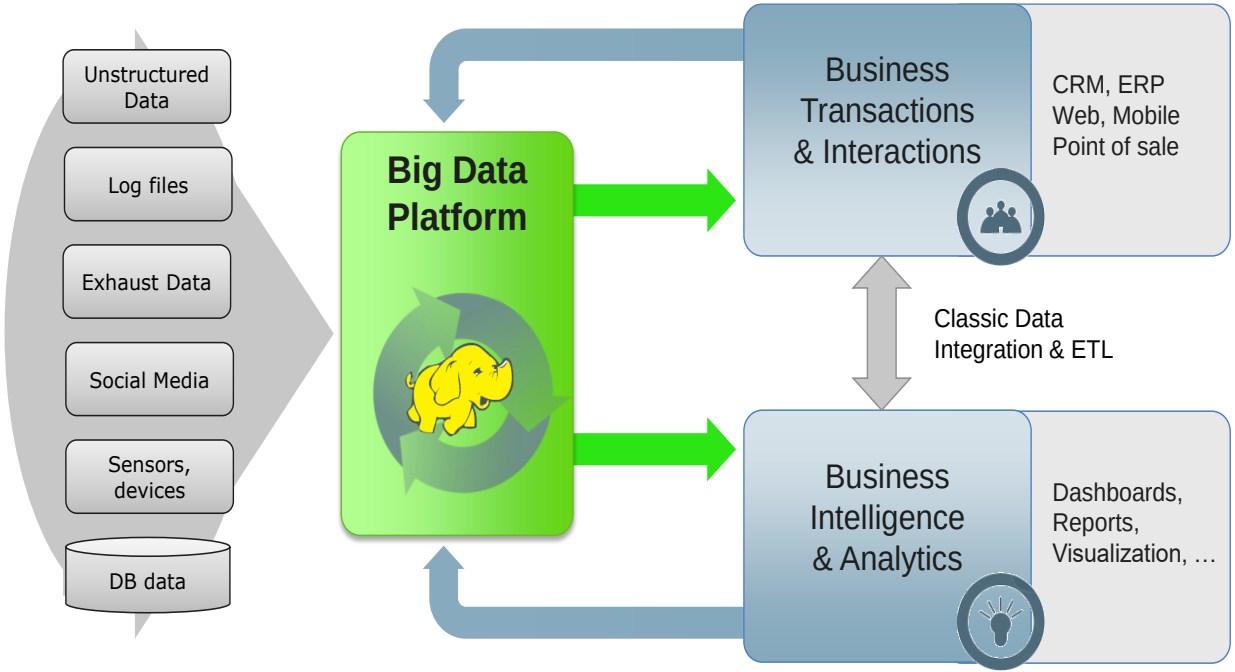
- Data access, cleansing and categorization
- Data storage and management



- De-centralized/multi-server architectures management
  - Performance bottlenecks, poor responsiveness, crash
  - Hardware requirements
  - (Big) Data management tools like Cluster mgt, multi-region deployment, etc.
- Non stable software distributions, rapid evolving fields
- Missing Skills

How to Implement Hadoop?

Typical Project Development Steps



1 Capture Big Data
Collect data from all sources structured &unstructured

2 Process
Transform, refine, aggregate, analyze, report

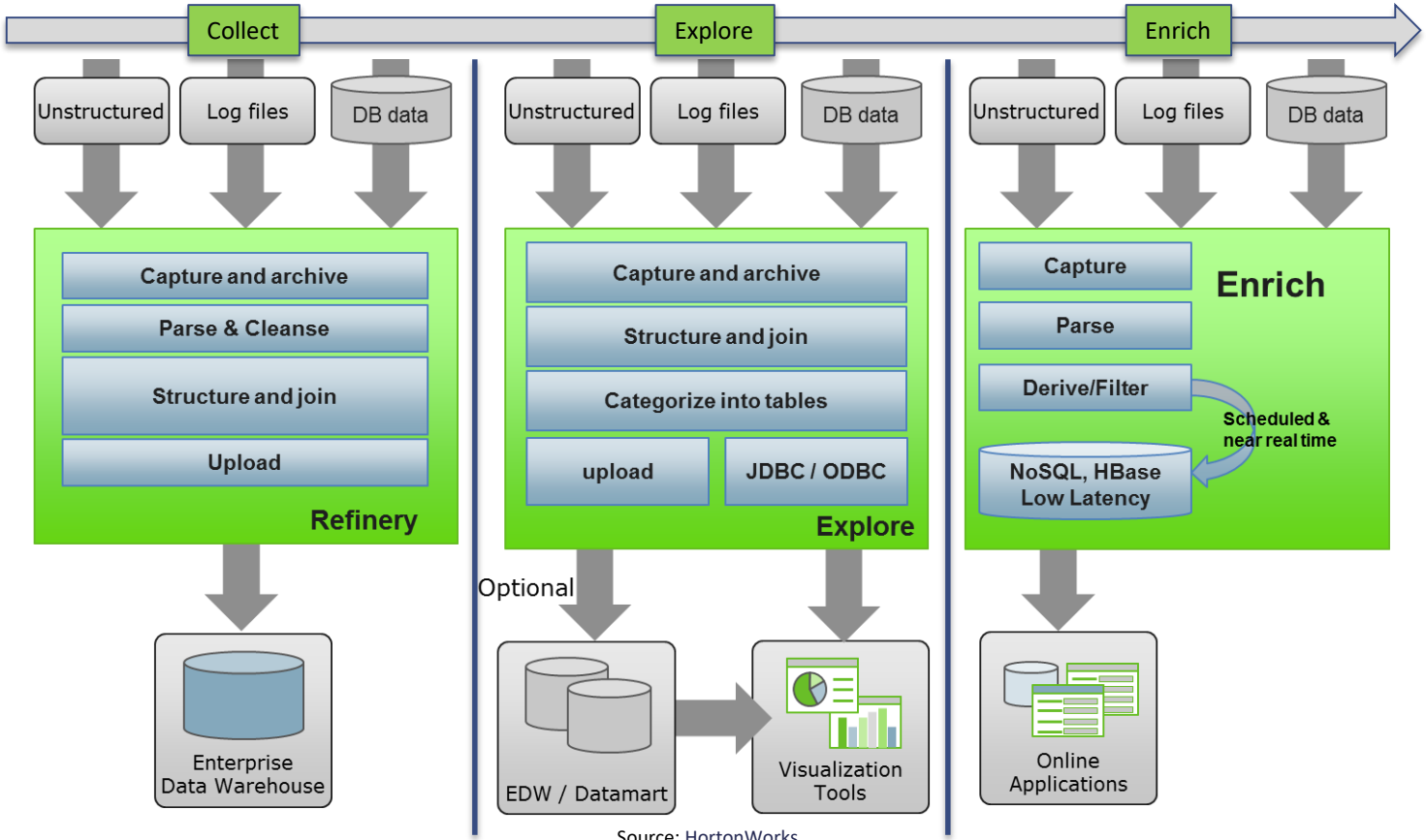
3 Distribute Results
Interoperate and share data with applications/analytics

4 Feedback
Use operational data w/in the big data platform

Source: HortonWorks

How to Implement Hadoop?

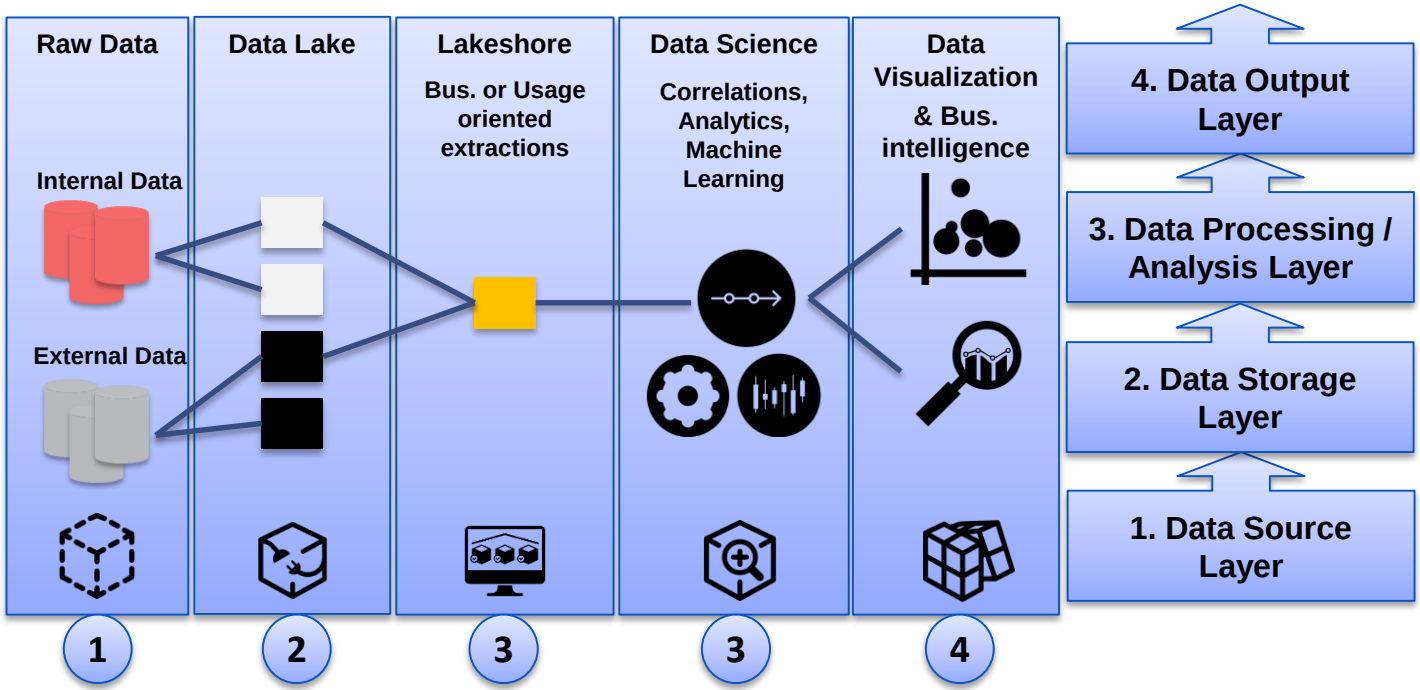
Typical Three Steps Process



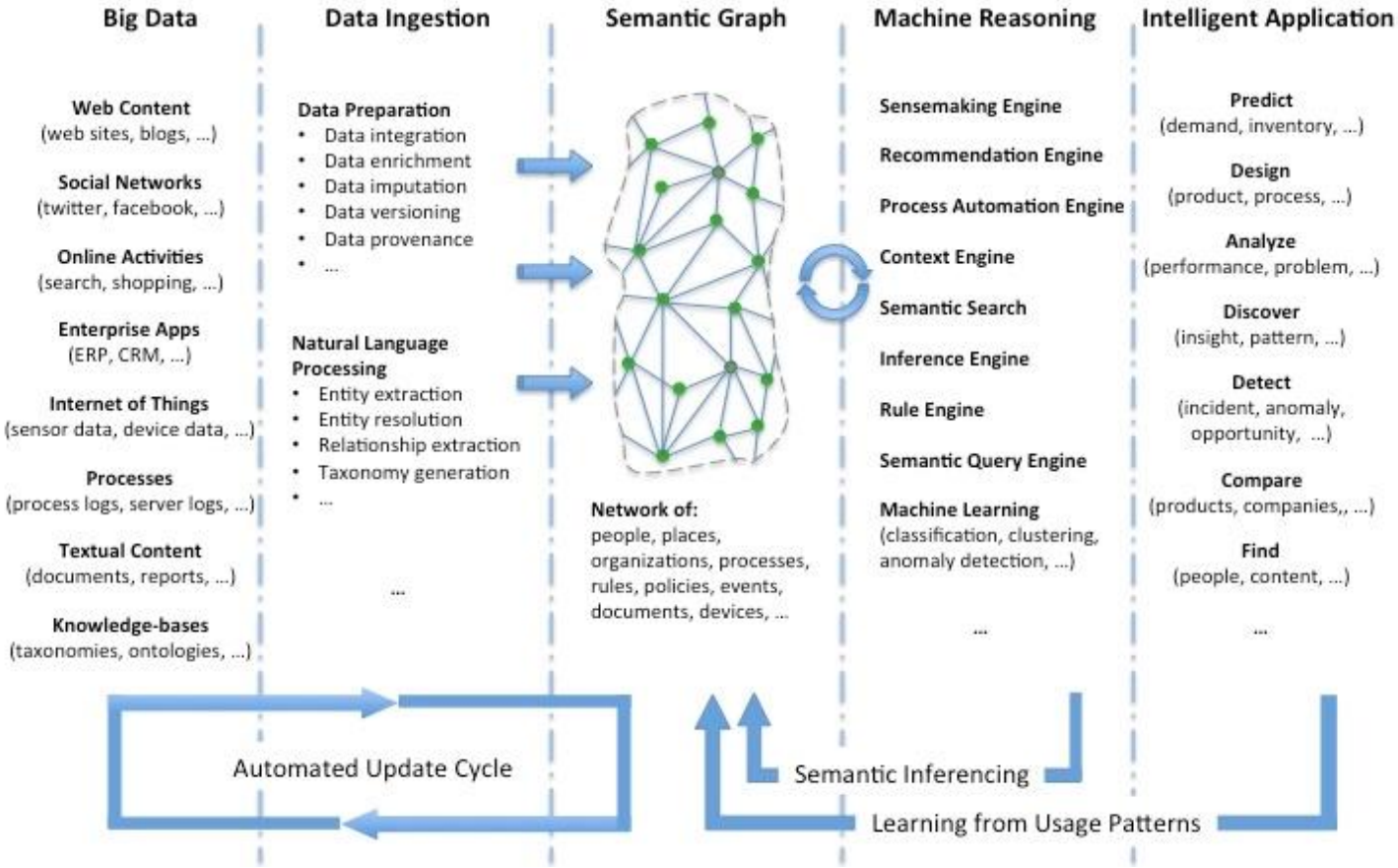
Source: HortonWorks

How to Implement Hadoop?

Typical Project Development Steps



How to Implement Hadoop?



Source: Cirrus Shakeri

How to Implement Hadoop?

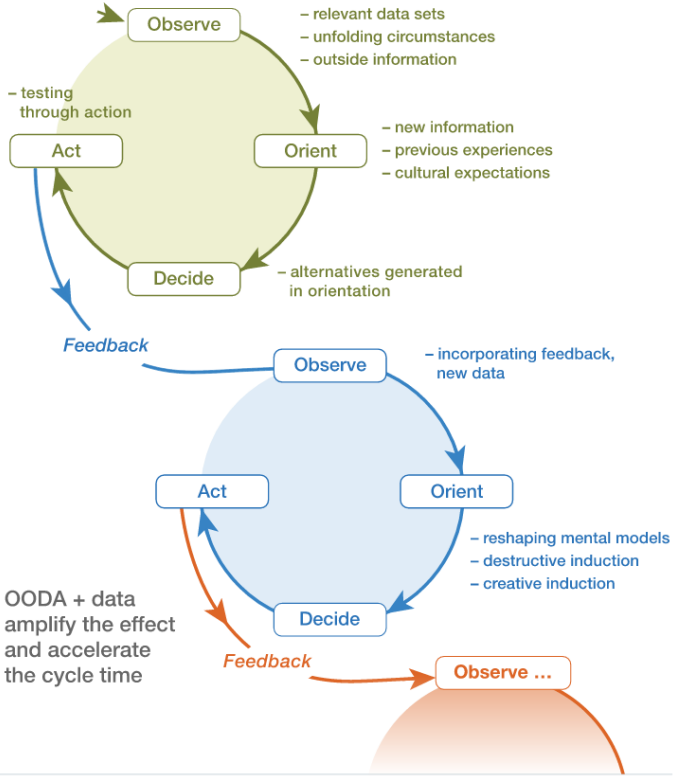
CRISP Methodology

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Integrate Data** *Merged Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Format Data *Reformatted Data* *Dataset* *Dataset Description*			

Source: The Modeling Agency and sv-europe

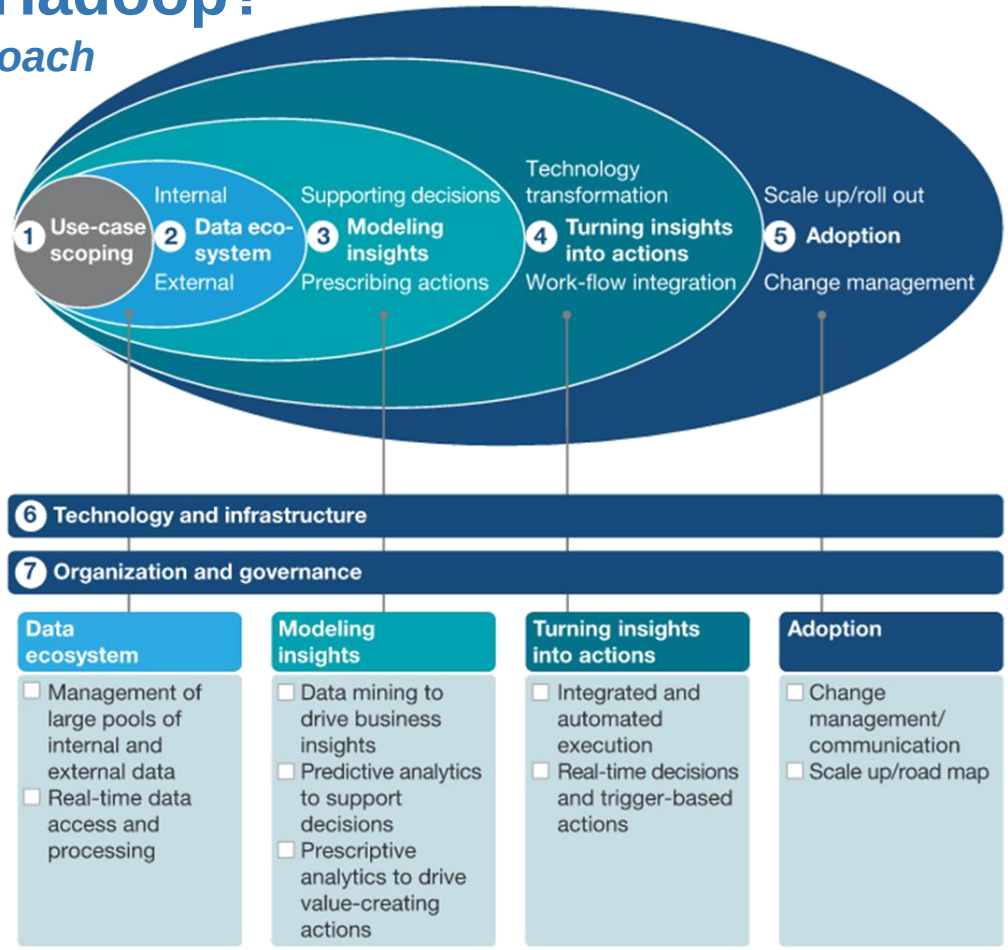
How to Implement Hadoop?

McKinsey Seven Steps Approach



[1]Observe, orient, decide, and act, a strategic decision-making model developed by US Air Force Colonel John R. Boyd.

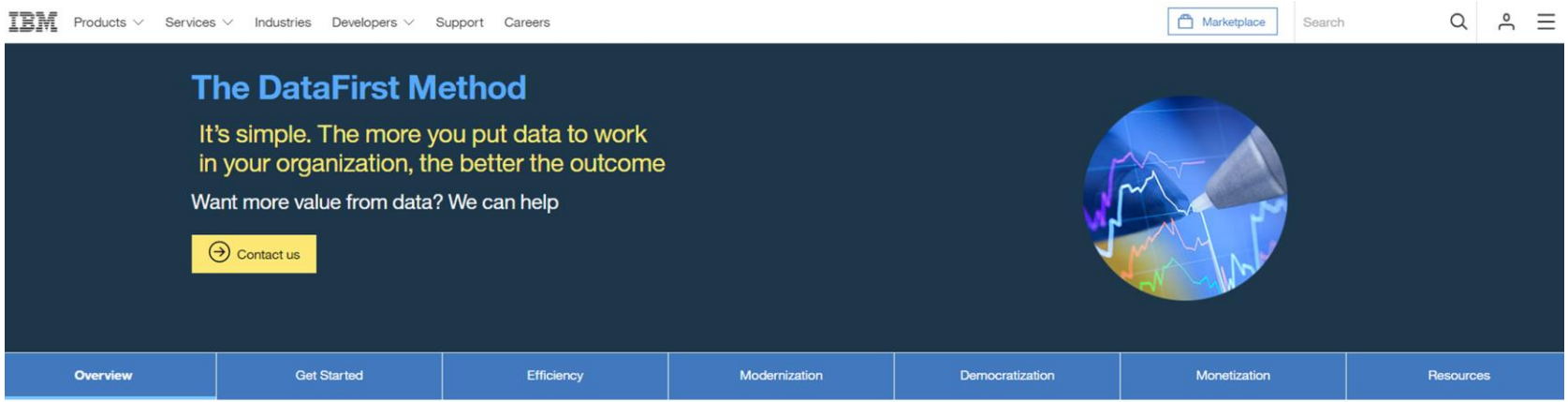
McKinsey&Company



Source: McKinsey

How to Implement Hadoop?

IBM DataFirst Method



What is The DataFirst Method?

Our methodology provides the strategy, expertise and game plan to help you get the most value from data. The DataFirst Method places the different uses of data into a maturity model so you can assess where you are today – and more importantly identify what's best to do next. It provides a set of workshops, methods and proven practices to ensure success on your journey to becoming a data-driven business.

Read data sheet (333 KB)

Start Anywhere

Focus on your biggest business opportunity.

Fill the Gaps

Strategy. Expertise. Skills. No more and no less.

Build Value at Every Step

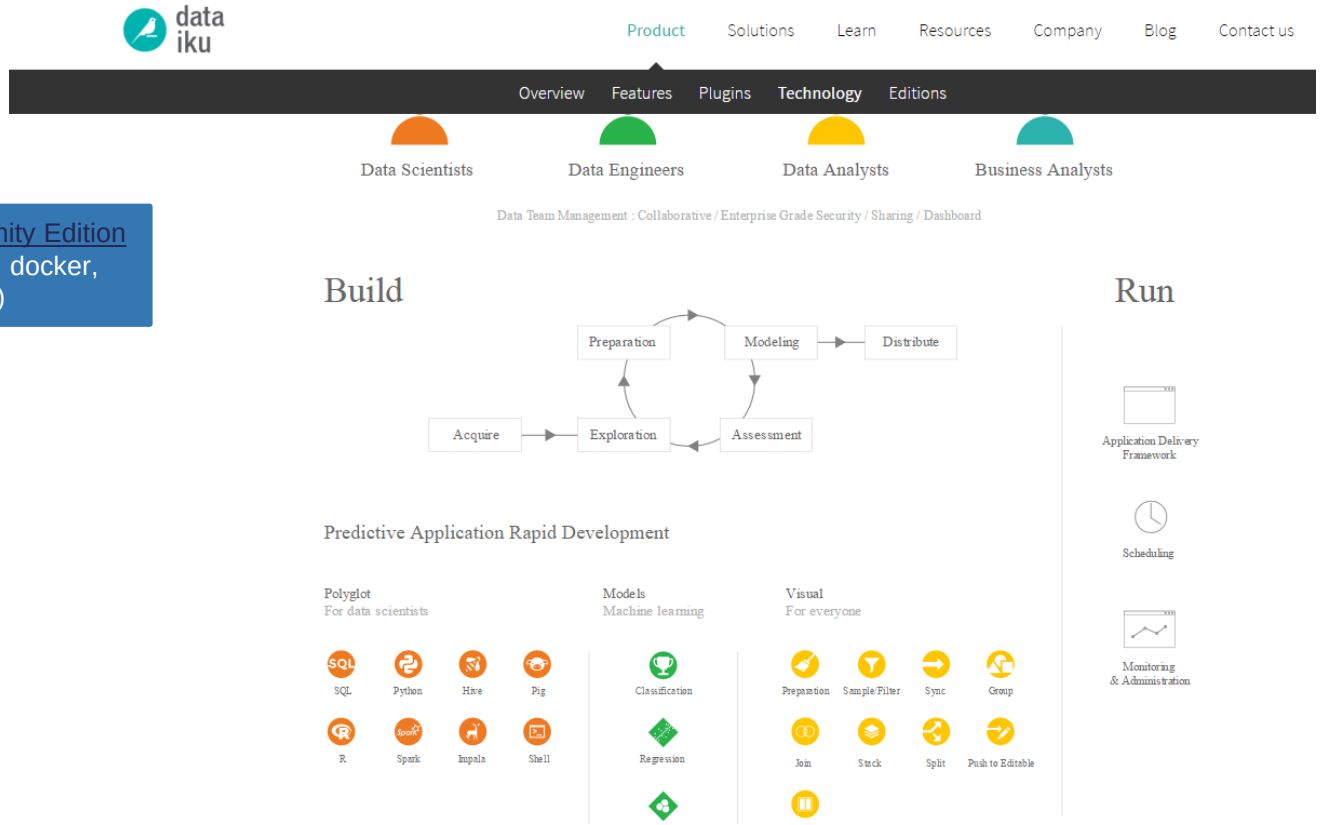
Achieve a data-driven culture, one initiative at a time.

Source: IBM

How to Implement Hadoop?

Dataiku Method supported by a tool

Free Community Edition (mac, linux, docker, aws)



http://www.dataiku.com/dss/

Thank You