ANALYTIXLABS

Introduction to BigData & Hadoop

Introduction to BigData

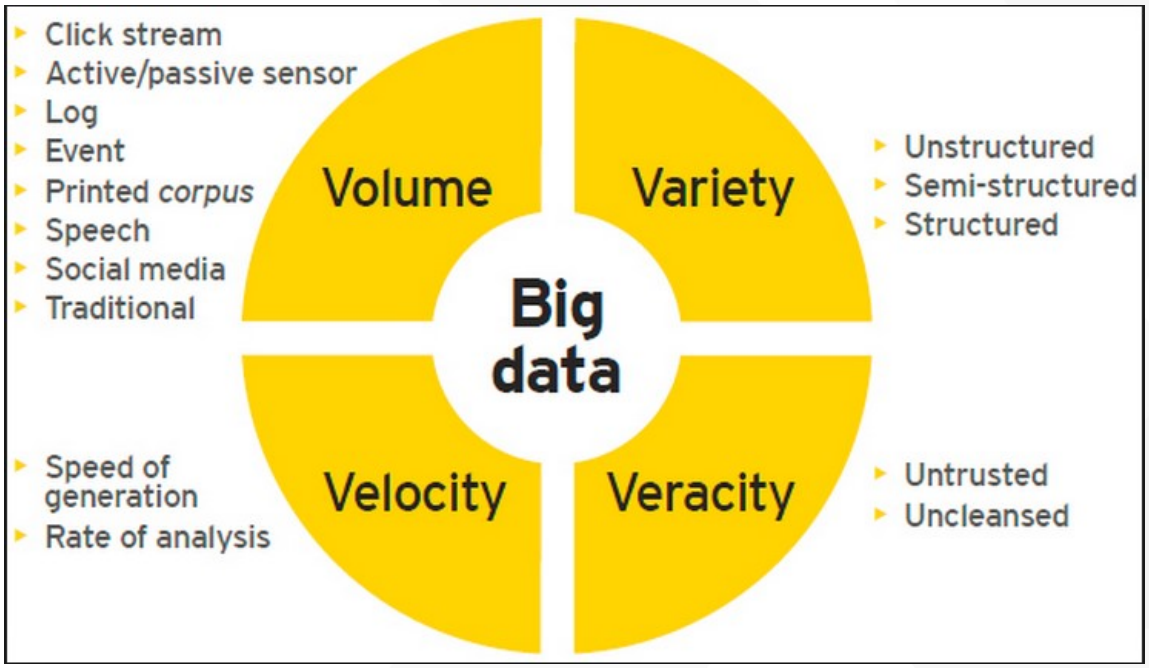
What is Big data?

- Is it technology?
- Is it solution?
- Is it problem?
- Is it platform?
- Is it statement/phrase?

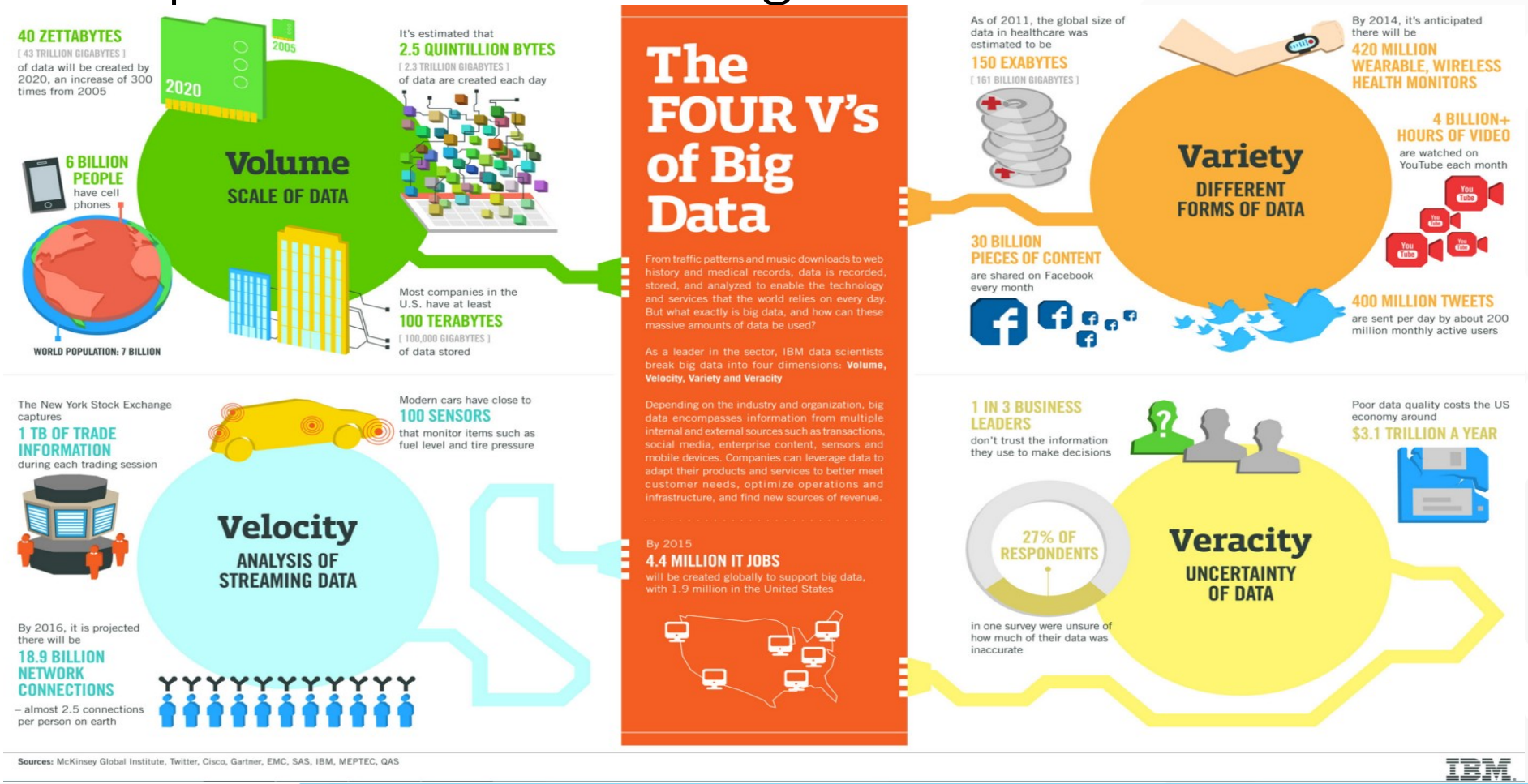
Big data – Myths(Perceptions)

- ✓ It’s Big : You need to have lots of data to talk about big data
- ✓ You need to apply it right away
- ✓ The more granular the data, the better
- ✓ Big Data is good data
- ✓ Big Data means that analysts become all-important
- ✓ Big Data gives you concrete answers
- ✓ Big Data predicts the future
- ✓ Big Data is a magical solution
- ✓ Big Data can create self-learning algorithms
- ✓ Big Data is only for big corporations
- ✓ We Have So Much Data, We Don't Need to Worry About Every Little Data Flaw
- ✓ Big Data Technology Will Eliminate the Need for Data Integration
- ✓ It's Pointless Using a Data Warehouse for Advanced Analytics
- ✓ Data Lakes Will Replace the Data Warehouse
- ✓ Hadoop is the holy grail of big data
- ✓ Machine Learning Overcomes Human Bias

Big Data – 4 V's


Click stream
Active/passive sensor
Log
Event
Printed corpus
Speech
Social media
Traditional
Volume
Variety
Unstructured
Semi-structured
Structured
Big data
Speed of generation
Rate of analysis
Velocity
Veracity
Untrusted
Uncleansed

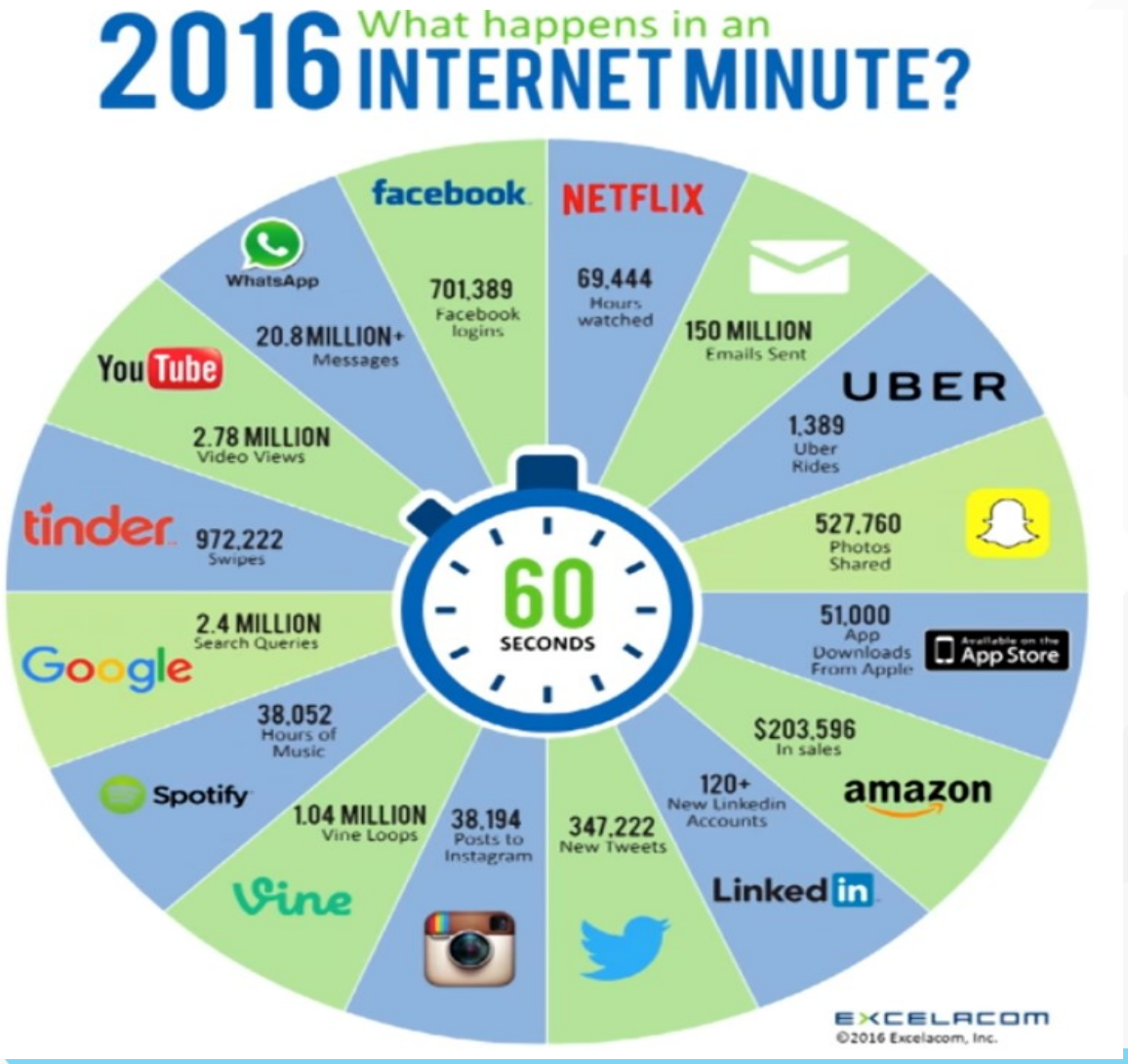
InfoGraphics: Definition of Big Data



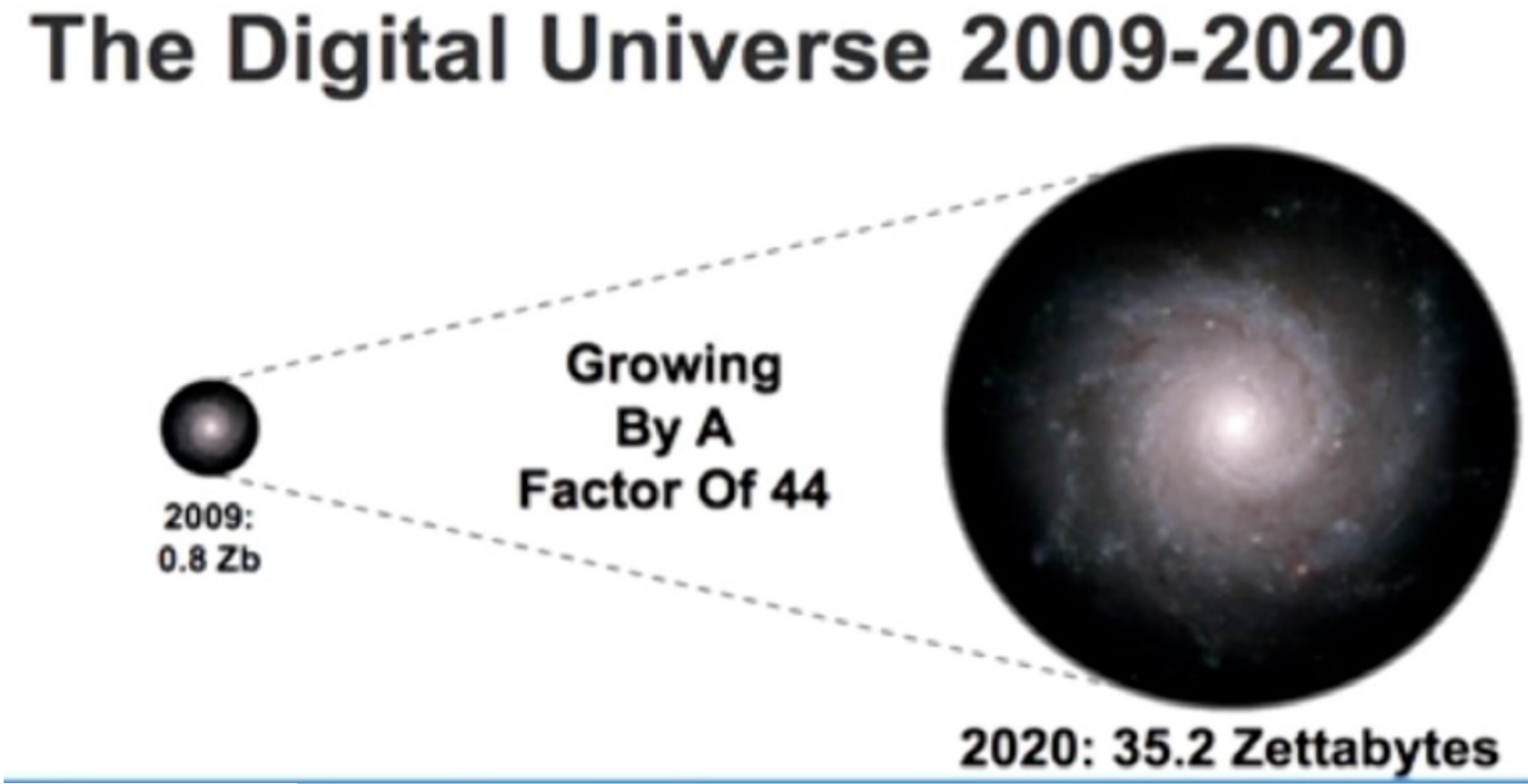
Big Data Sources

- Sample list of Sources

1. eCommerece Logs
2. Weather Data
3. Social media data
4. Micro blogging data
5. Product Catalogs
6. Unstructured text
7. Video Streams
8. Application Logs
9. Web access logs
10. Web Proxy Logs
8. Call Detail Records (CDRs)
9. Click Stream Data
10. Message queues
11. Packet data
12. Configuration files
13. DB audit logs
14. Management & Logging APIs
15. OS Metrics, Status and diagnostic command outputs
16. Syslog
17. Logs from DNS and DHCP


2016 What happens in an INTERNET MINUTE?
facebook
701,389
Facebook logins
NETFLIX
69,444
Hours watched
WhatsApp
20.8 MILLION+
Messages
150 MILLION
Emails Sent
UBER
1,389
Uber Rides
YouTube
2.78 MILLION
Video Views
527,760
Photos Shared
tinder
972,222
Swipes
60
SECONDS
51,000
App Downloads From Apple
Available on the App Store
2.4 MILLION
Search Queries
Google
$203,596
In sales
38,052
Hours of Music
amazon
120+
New Linkedin Accounts
Spotify
1.04 MILLION
Vine Loops
38,194
Posts to Instagram
347,222
New Tweets
LinkedIn
Vine
EXCELACOM
©2016 Excelacom, Inc.

Growth of Digital data



Ref: IDC

What is the challenge?

- The challenges include **capture, curation, storage, search, sharing transfer, analysis and visualization**

- The main challenge lies in identifying the value, the relevant information within this data, and then transforming and extracting that data for further analysis.

Data Deluge

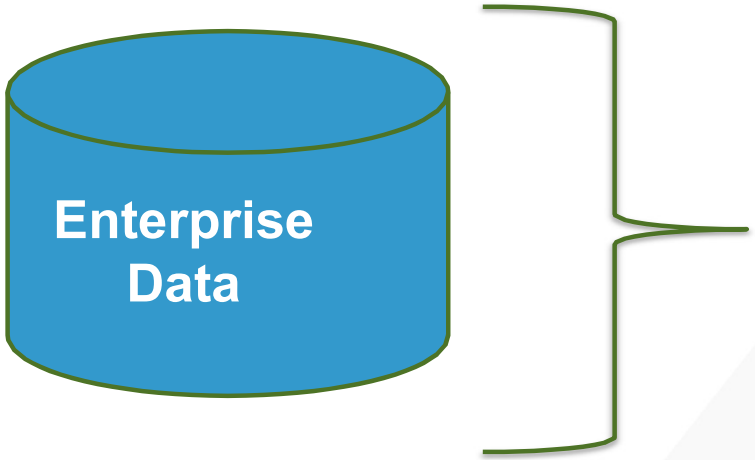
"We are drowning in information and starving for knowledge"

– John Naisbitt

Source: Megatrends, 1982

Quiz

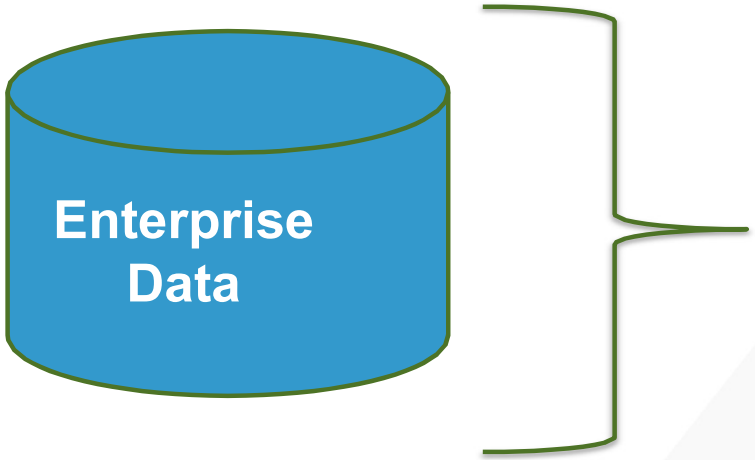
What percentage of enterprise data do firms use for analytics?



A. 12%
B. 34%
C. 53%
D. 76%

Quiz

What percentage of enterprise data do firms use for analytics?



A. 12%
B. 34%
C. 53%
D. 76%

Big Data- Scenarios

- ✓ Web and e-tailing
  - ✓ Recommendation Engines
  - ✓ Ad Targeting
  - ✓ Search Quality
  - ✓ Abuse and Click Fraud Dete

- ✓ Government
  - ✓ Fraud Detection And Cyber Security
  - ✓ Welfare schemes
  - ✓ Justice

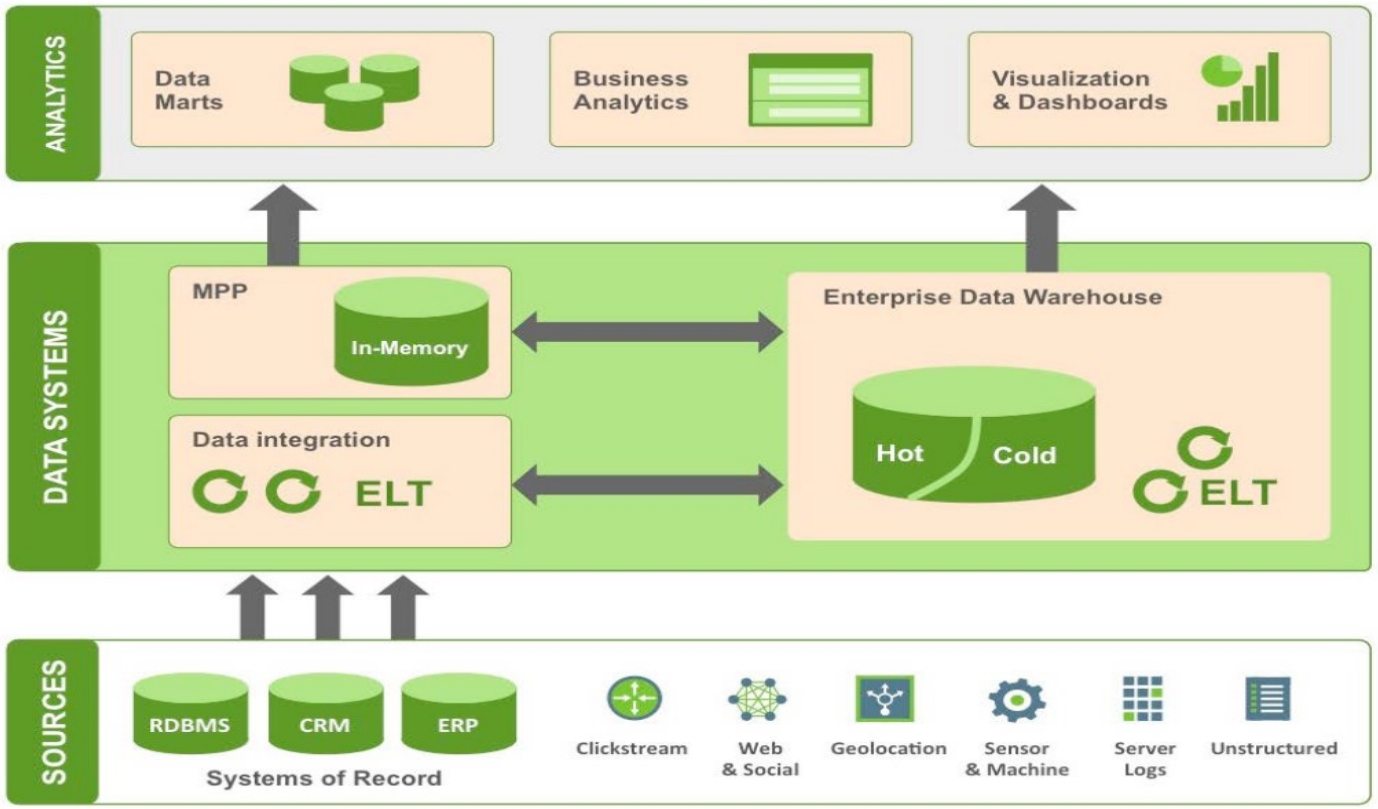
- ✓ Healthcare & Life Sciences
  - ✓ Health information exchange
  - ✓ Gene sequencing
  - ✓ Serialization
  - ✓ Healthcare service quality improvements
  - ✓ Drug Safety

- ✓ Telecommunications
  - ✓ Customer Churn Prevention
  - ✓ Network Performance Optimization
  - ✓ Calling Data Record (CDR) Analysis
  - ✓ Analyzing Network to Predic Failure

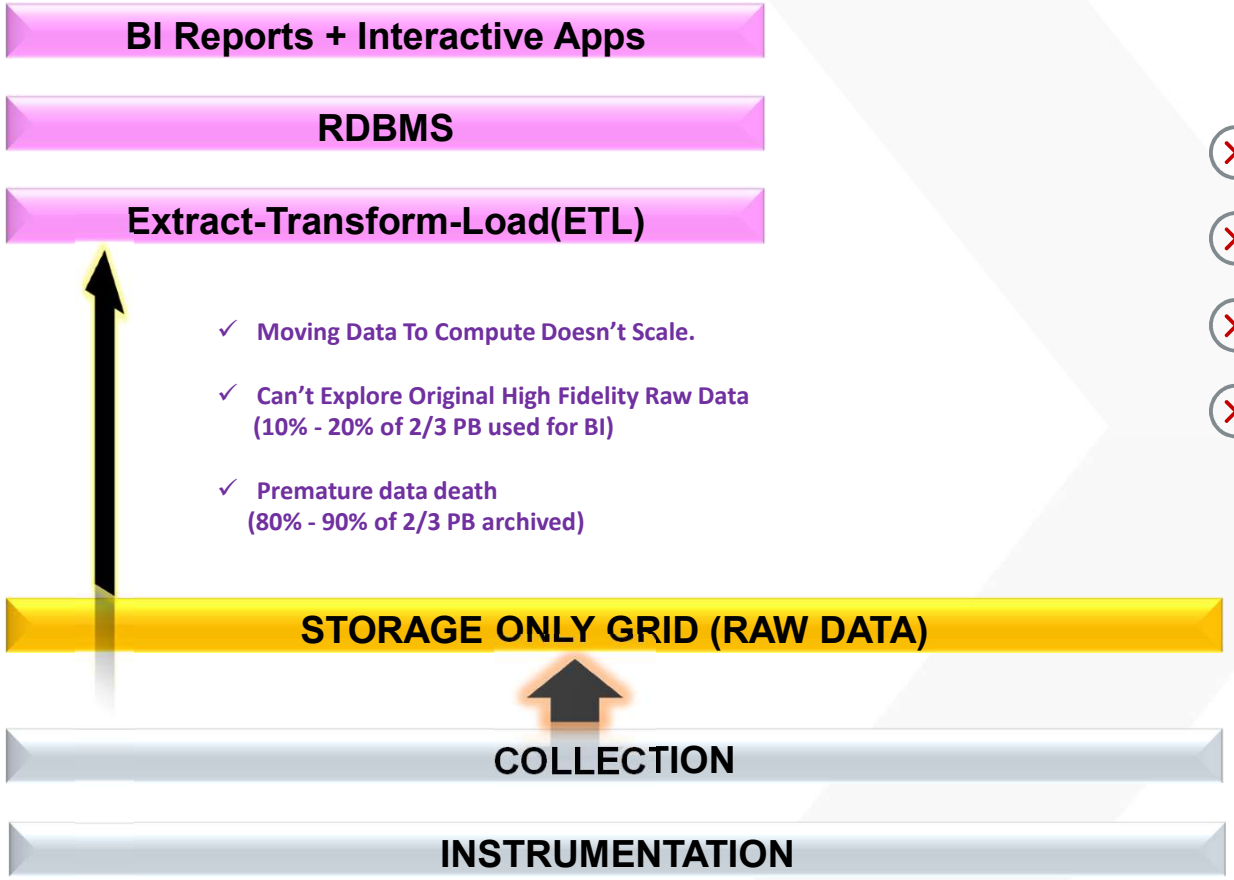
- ✓ Banks and Financial services
  - ✓ Modeling True Risk
  - ✓ Threat Analysis
  - ✓ Fraud Detection
  - ✓ Trade Surveillance
  - ✓ Credit Scoring And Analysis

- ✓ Retail
  - ✓ Point of sales Transaction Analysis
  - ✓ Customer Churn Analysis
  - ✓ Sentiment Analysis

Existing Data Architecture

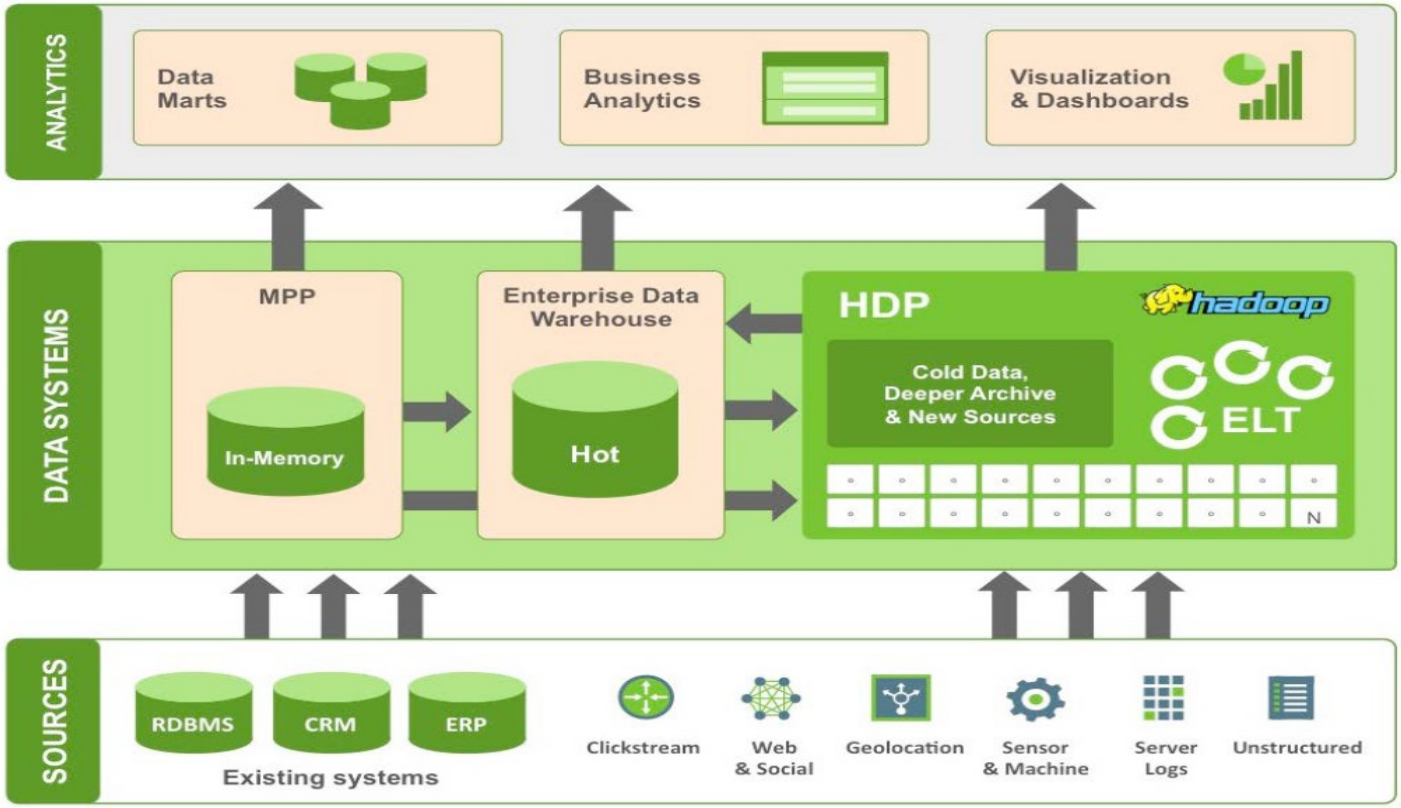


Limitations of Existing Data Analytics Architecture

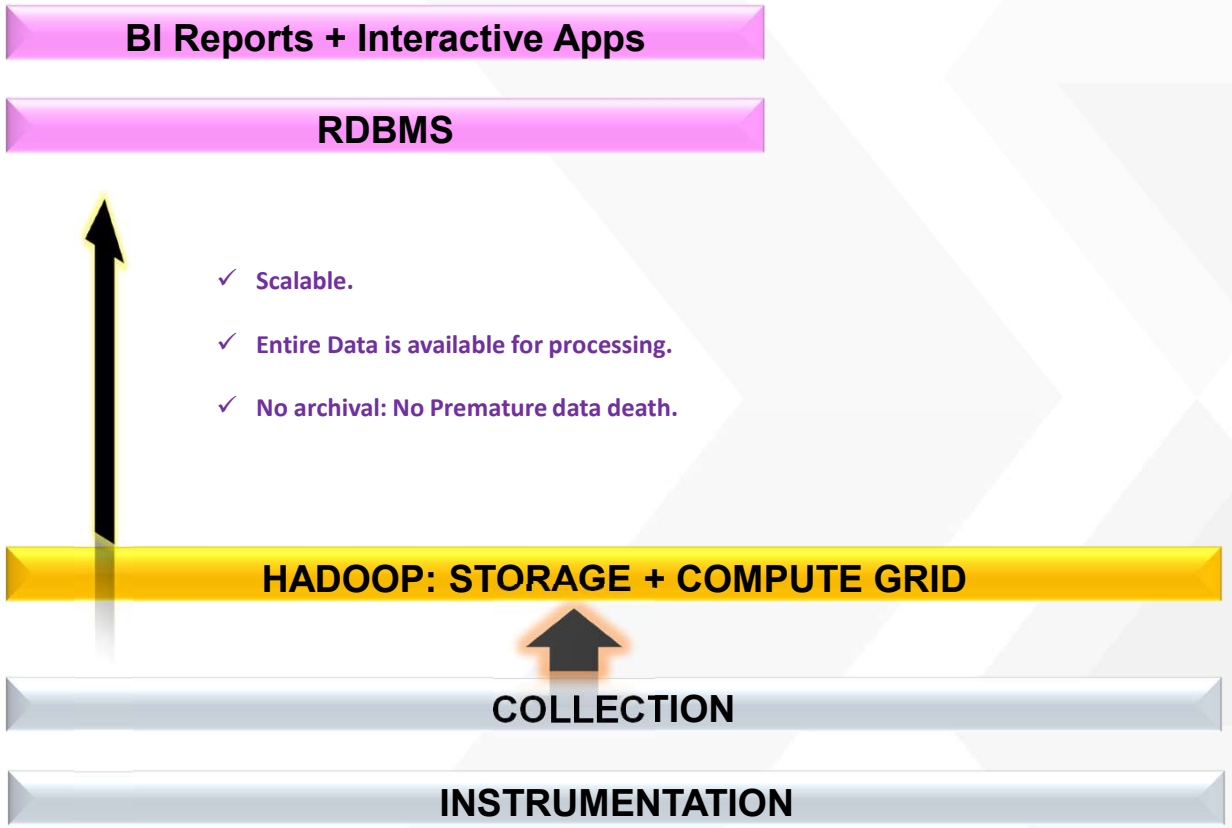


- Constrains data to specific apps
- No insight across ALL data
- Built for structured data
- Does not scale (cost and tech)

An Emerging Data Architecture



Emerging Data Analytics Architecture



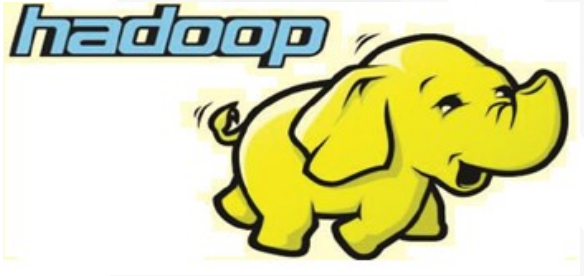

HADOOP

What is Apache Hadoop?

- Open source software framework designed for storage and processing of large scale data on clusters of commodity hardware

- Created by Doug Cutting and Mike Carafella in 2005.

- Cutting named the program after his son's toy elephant.



Motivations for Hadoop

- What considerations led to its design

Motivations for Hadoop

- What were the limitations of earlier large-scale computing?
- What requirements should an alternative approach have?
- How does Hadoop address those requirements?

Early Large Scale Computing

- Historically computation was processor-bound
  - Data volume has been relatively small
  - Complicated computations are performed on that data
- Advances in computer technology has historically centered around improving the power of a single machine

Distributed Systems

- Allows developers to use multiple machines for a single task

Distributed System: Problems

- Programming on a distributed system is much more complex
  - Synchronizing data exchanges
  - Managing a finite bandwidth
  - Controlling computation timing is complicated

Distributed System: Problem

"You know you have a distributed system when the crash of a computer you've never heard of stops you from getting any work done." – Leslie Lamport

- Distributed systems designed with the expectation of failure

Nice blog on Fallacies of distributed computing

https://blog.newrelic.com/2011/01/06/the-fallacies-of-distributed-computing-reborn-the-cloud-era/

Distributed System: Data Storage

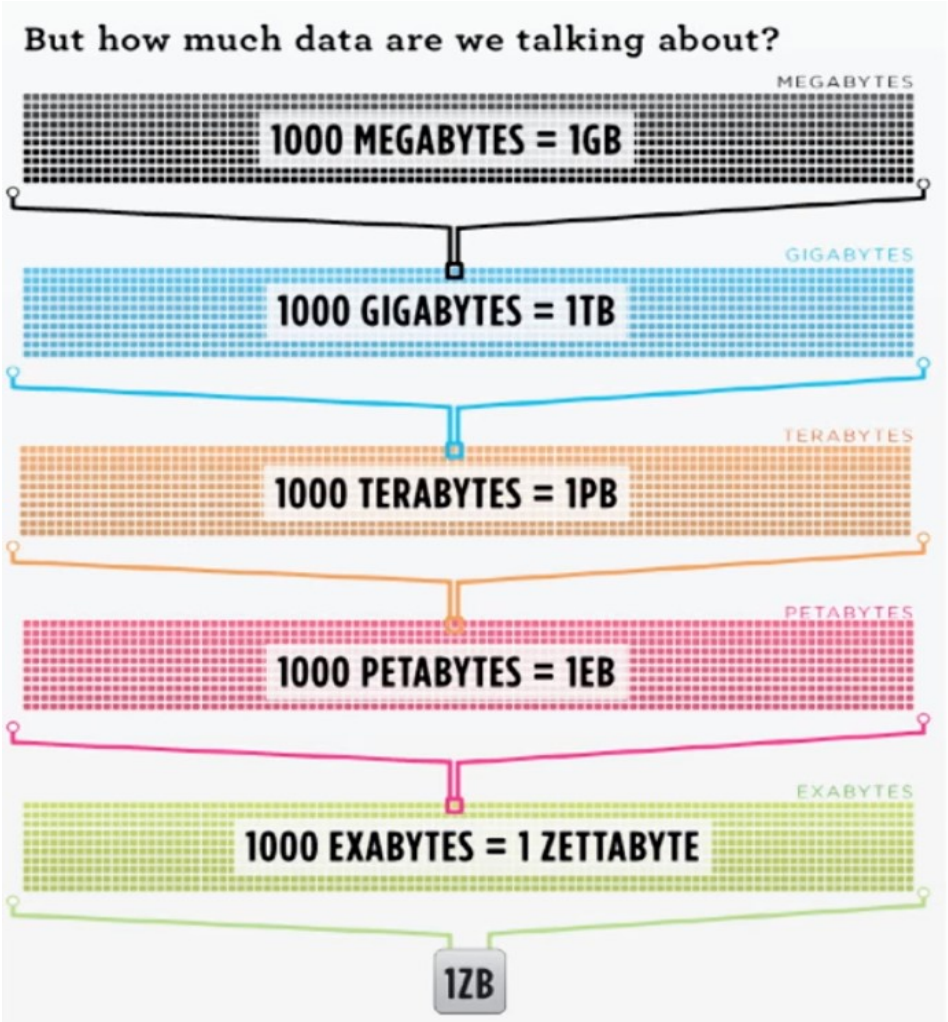
- Typically divided into Data Nodes and Compute Nodes
- At compute time, data is copied to the Compute Nodes
- Fine for relatively small amounts of data

- Modern systems deal with far more data than was gathering in the past

How much data?

- Facebook
  - 500 TB per day
- Yahoo
  - Over 170 PB
- eBay
  - Over 6 PB

- Getting the data to the processors becomes the bottleneck


But how much data are we talking about?
MEGABYTES
1000 MEGABYTES = 1GB
GIGABYTES
1000 GIGABYTES = 1TB
TERABYTES
1000 TERABYTES = 1PB
PETABYTES
1000 PETABYTES = 1EB
EXABYTES
1000 EXABYTES = 1 ZETTABYTE
1ZB

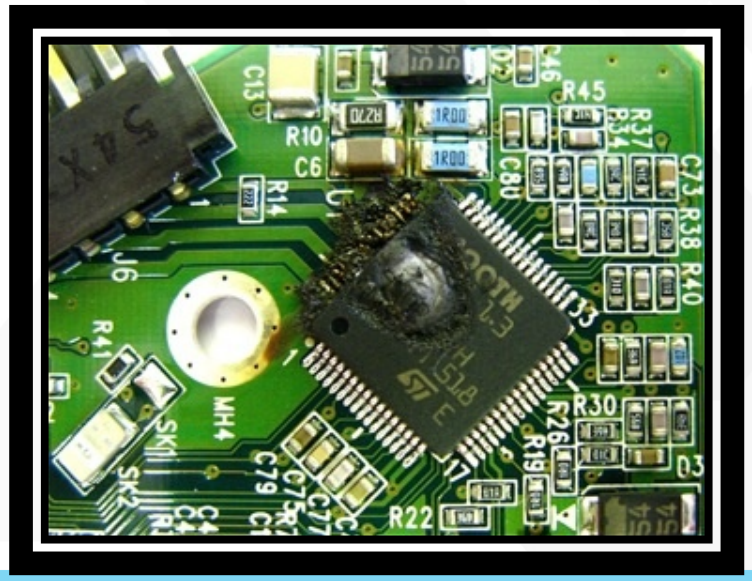
Requirements for Hadoop

- Must support partial failure
- Must be scalable

Partial Failures

- Failure of a single component must not cause the failure of the entire system only a degradation of the application performance

  - Failure should not result in the loss of any data



Component Recovery

- If a component fails, it should be able to recover without restarting the entire system
- Component failure or recovery during a job must not affect the final output

Scalability

- Increasing resources should increase load capacity
- Increasing the load on the system should result in a graceful decline in performance for all jobs
  - Not system failure

Hadoop

- Based on work done by Google in the early 2000s
  - “The Google File System” in 2003
  - “MapReduce: Simplified Data Processing on Large Clusters” in 2004
- The core idea was to distribute the data as it is initially stored
  - Each node can then perform computation on the data it stores without moving the data for the initial processing

Core Hadoop Concepts

- Applications are written in a high-level programming language
  - No network programming or temporal dependency
- Nodes should communicate as little as possible
  - A “shared nothing” architecture
- Data is spread among the machines in advance
  - Perform computation where the data is already stored as often as possible

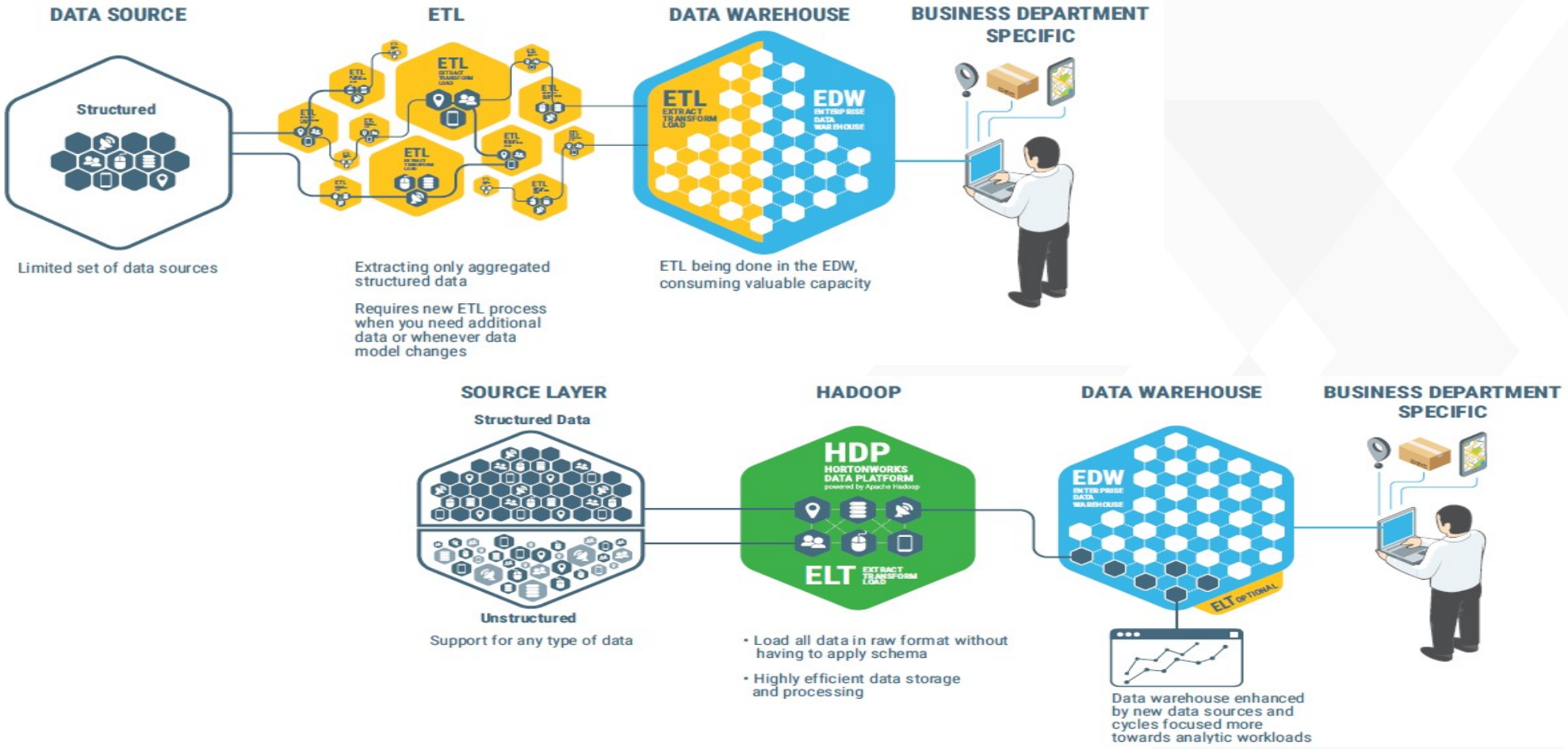
High-Level Overview

- When data is loaded onto the system it is divided into blocks
  - Typically 64MB or 128MB
- Tasks are divided into two phases
  - Map tasks which are done on small portions of data where the data is stored
  - Reduce tasks which combine data to produce the final output
- A master program allocates work to individual nodes

Fault Tolerance

- Failures are detected by the master program which reassigns the work to a different node
- Restarting a task does not affect the nodes working on other portions of the data
- If a failed node restarts, it is added back to the system and assigned new tasks
- The master can redundantly execute the same task to avoid slow running nodes

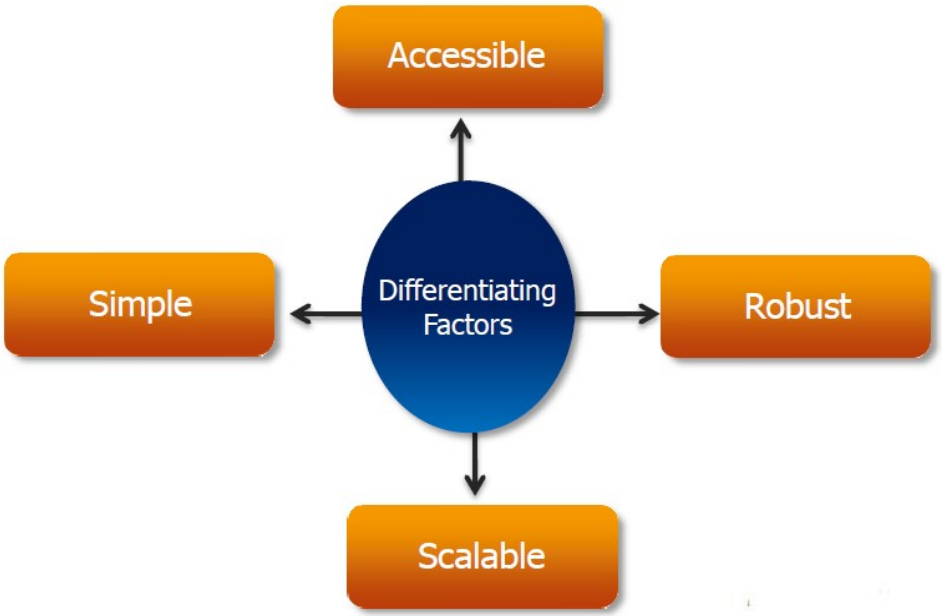
DBMS vs. HADOOP

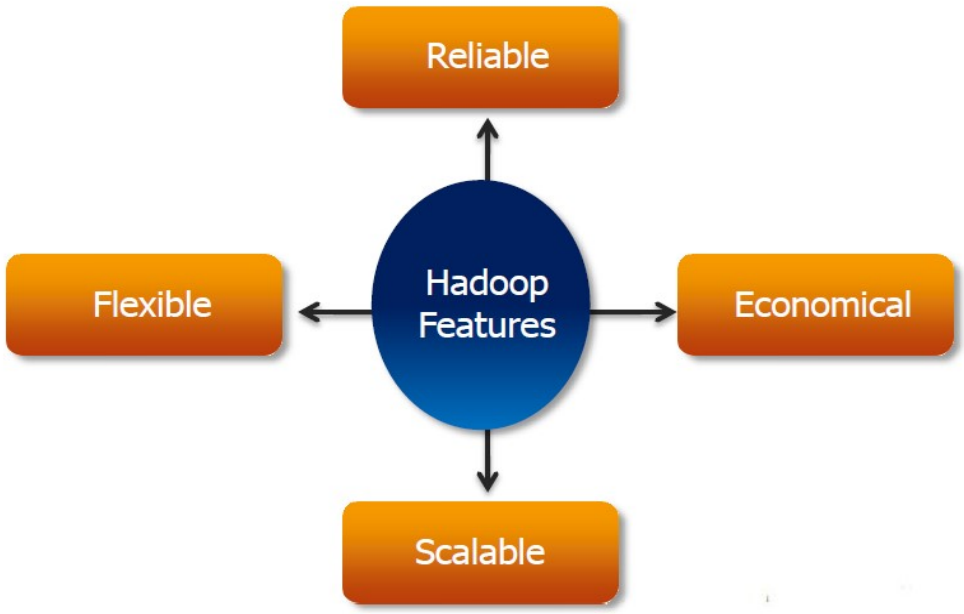


DBMS vs. HADOOP

DBMS	Hadoop
▪ Data is kept in relational database system	▪ Data is kept in flat file structure
▪ Can store and process only structured data (i.e. tabular format)	▪ Can store and process multi-structured and unstructured data
▪ Known Entity/ resources	▪ Growing, Complexities, Wide etc.
▪ Schema required on Write	▪ Schema required on Read
▪ Can be faster than Hadoop for simple query	▪ Outperforms database for complex queries
▪ Standardized language supported by many vendors	▪ Immature technology so significant changes are expected

Why Hadoop?


Accessible
Simple
Differentiating Factors
Robust
Scalable


Reliable
Flexible
Hadoop Features
Economical
Scalable

Why Hadoop?

- ✓ Supports use of inexpensive, commodity hardware
  - ✓ No RAID needed. Also, the servers need not be the latest and greatest hardware.
- ✓ Provides for simple, massive parallelism
- ✓ Provides resilience by replicating data and eliminating tape backups
- ✓ Provides locality of execution, as it knows where the data is
- ✓ Software free
- ✓ High quality support available at modest cost
- ✓ Certification available
- ✓ Easy to support when using GUI such as Cloudera Manager or Ambari
- ✓ Add-on tools available at relatively low cost, or in some cases no cost
- ✓ Evolving technology with a high degree of interest around the world

Hadoop tidbits

- ✓ Yahoo and Facebook are two well known companies that use Hadoop
- ✓ Yahoo has over 42,000 nodes
- ✓ These large installations don't use screws to mount the nodes in the rack, they use hook-and-loop (Velcro)
- ✓ MTBF about 1,000 days, so with 42,000 nodes about 42 fail each day
  - ✓ But it keeps running
- ✓ Largest amount of disk storage in use: Facebook over 100 Petabytes
  - ✓ Growing at over ½ Petabyte per day

- http://www.hadoopwizard.com/which-big-data-company-has-the-worlds-biggest-hadoop-cluster/
- http://gigaom.com/2012/06/13/how-facebook-keeps-100-petabytes-of-hadoop-data-online/
- http://hortonworks.com/hdp/downloads/
- http://www.cloudera.com
- https://wiki.apache.org/hadoop/PoweredBy
- http://www.datameer.com/product/
- http://alpinenow.com/
- http://spark.apache.org/docs/latest/quick-start.html

Hadoop Ecosystem

Hadoop Ecosystem

ETL

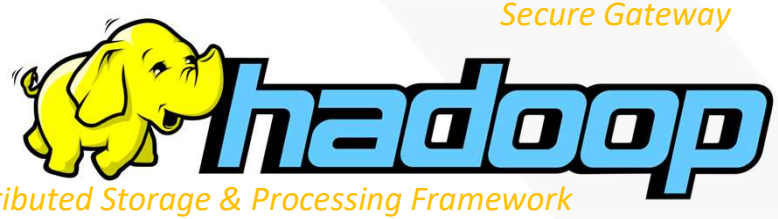
RDBMS Import/Export

Apache Ambari
http://incubator.apache.org/ambari

Cluster System Operations

Secure Gateway

Streaming Data Ingestion



Distributed Storage & Processing Framework

Workflow Management

SQL

Data Management

NoSQL DB

Search & Indexing

SQL on HBase

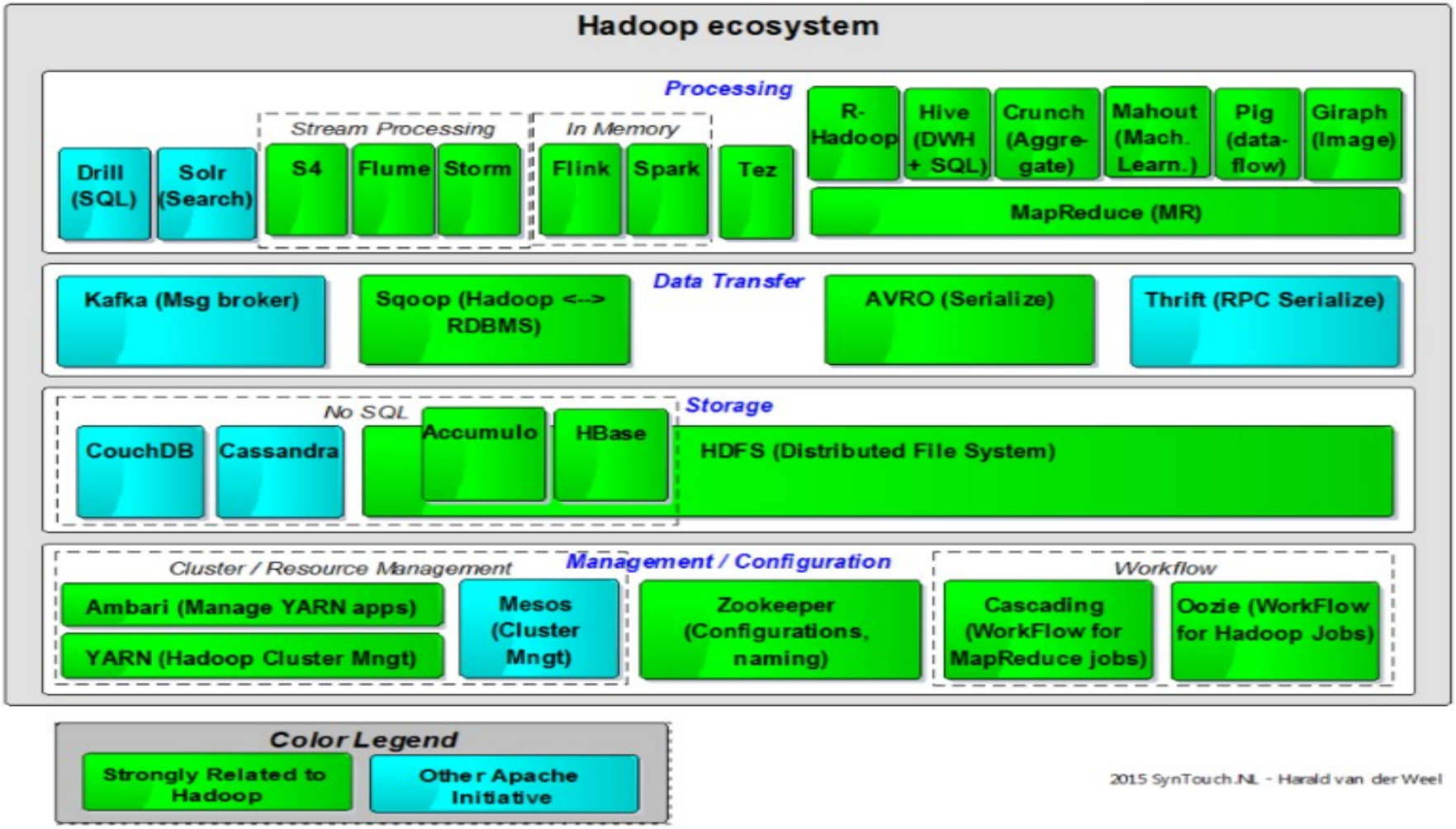
Distributed Registry

Even Faster Data Processing

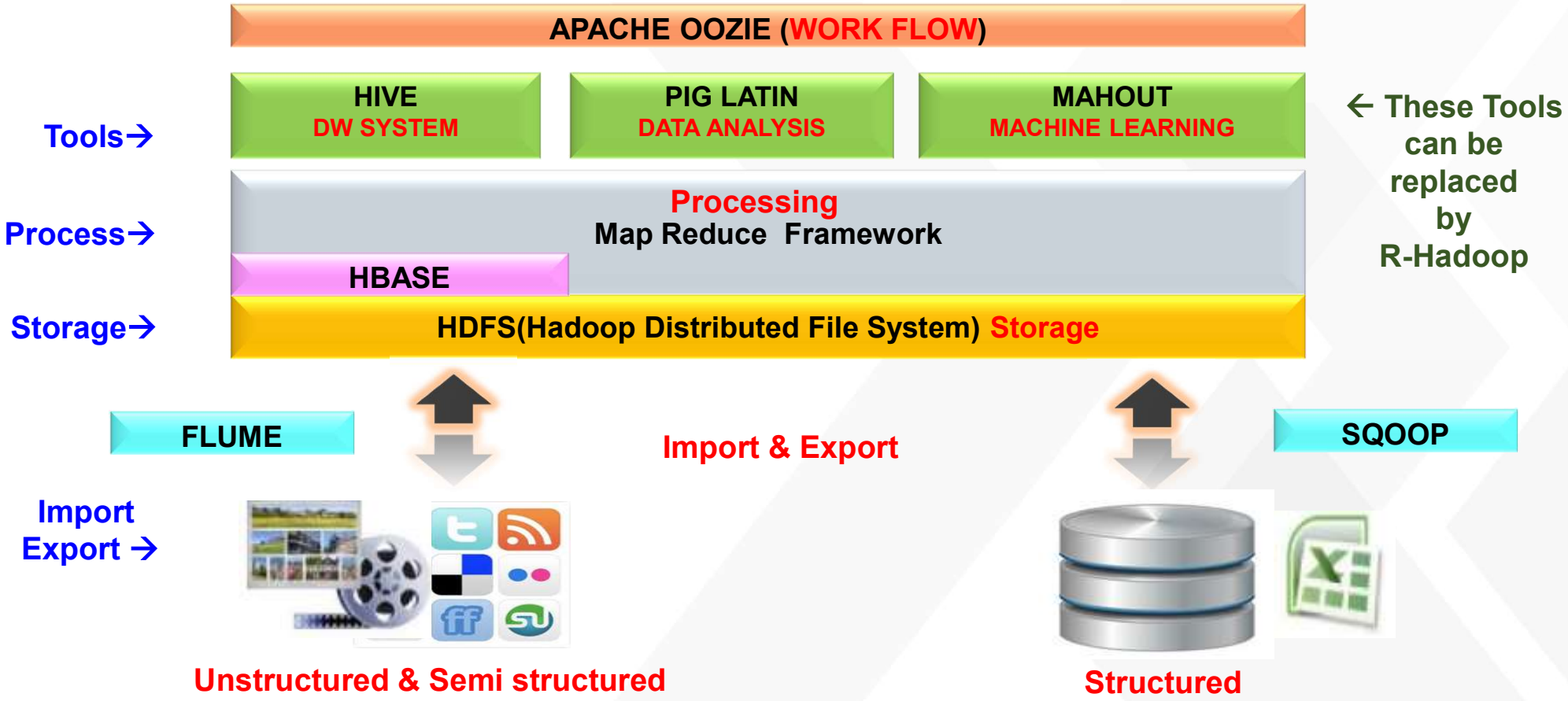
Machine Learning

Secure NoSQL DB

runs on

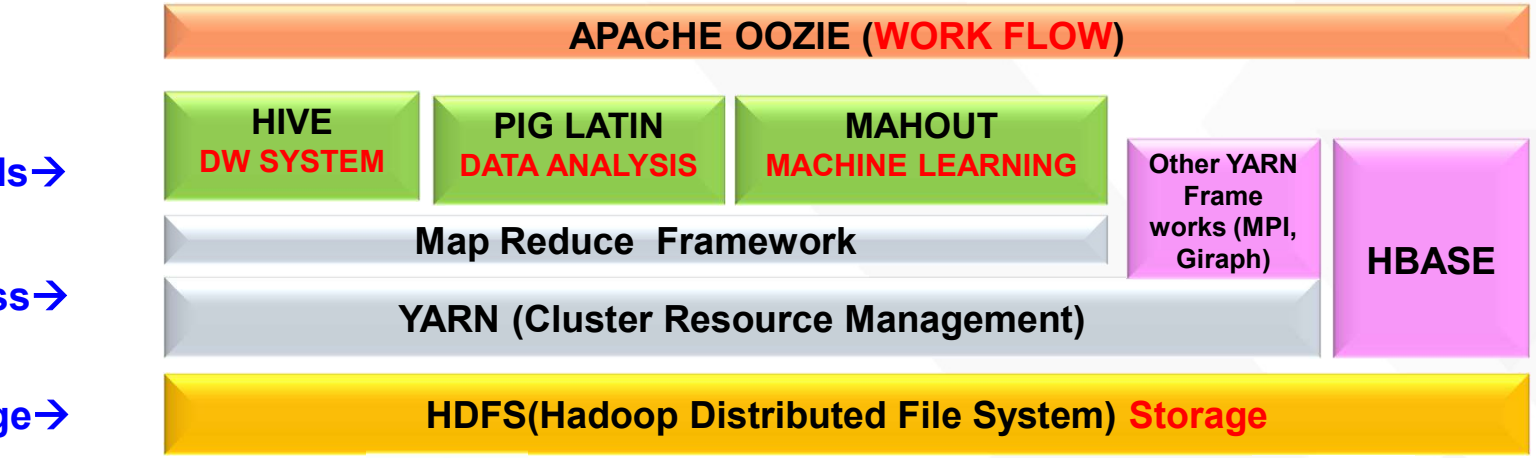

Hadoop ecosystem
Processing
Stream Processing
In Memory
Drill (SQL)
Solr (Search)
S4
Flume
Storm
Flink
Spark
Tez
R-Hadoop
Hive (DWH + SQL)
Crunch (Aggre-gate)
Mahout (Mach. Learn.)
Pig (data-flow)
Giraph (Image)
MapReduce (MR)
Data Transfer
Kafka (Msg broker)
Sqoop (Hadoop <--> RDBMS)
AVRO (Serialize)
Thrift (RPC Serialize)
Storage
No SQL
CouchDB
Cassandra
Accumulo
HBase
HDFS (Distributed File System)
Cluster / Resource Management
Management / Configuration
Workflow
Ambari (Manage YARN apps)
YARN (Hadoop Cluster Mngt)
Mesos (Cluster Mngt)
Zookeeper (Configurations, naming)
Cascading (WorkFlow for MapReduce jobs)
Oozie (WorkFlow for Hadoop Jobs)
Color Legend
Strongly Related to Hadoop
Other Apache Initiative
2015 SynTouch.NL - Harald van der Weel

Hadoop Eco-System – Analytics mapping – Hadoop 1.x



✓ Zookeeper is software project, providing an open source distributed configuration service and synchronization service and naming registry for large distributed system

Hadoop Eco-System – Analytics mapping – Hadoop 2.x



← These Tools can be replaced by R-Hadoop

FLUME

Import & Export

SQOOP

Import
Export →

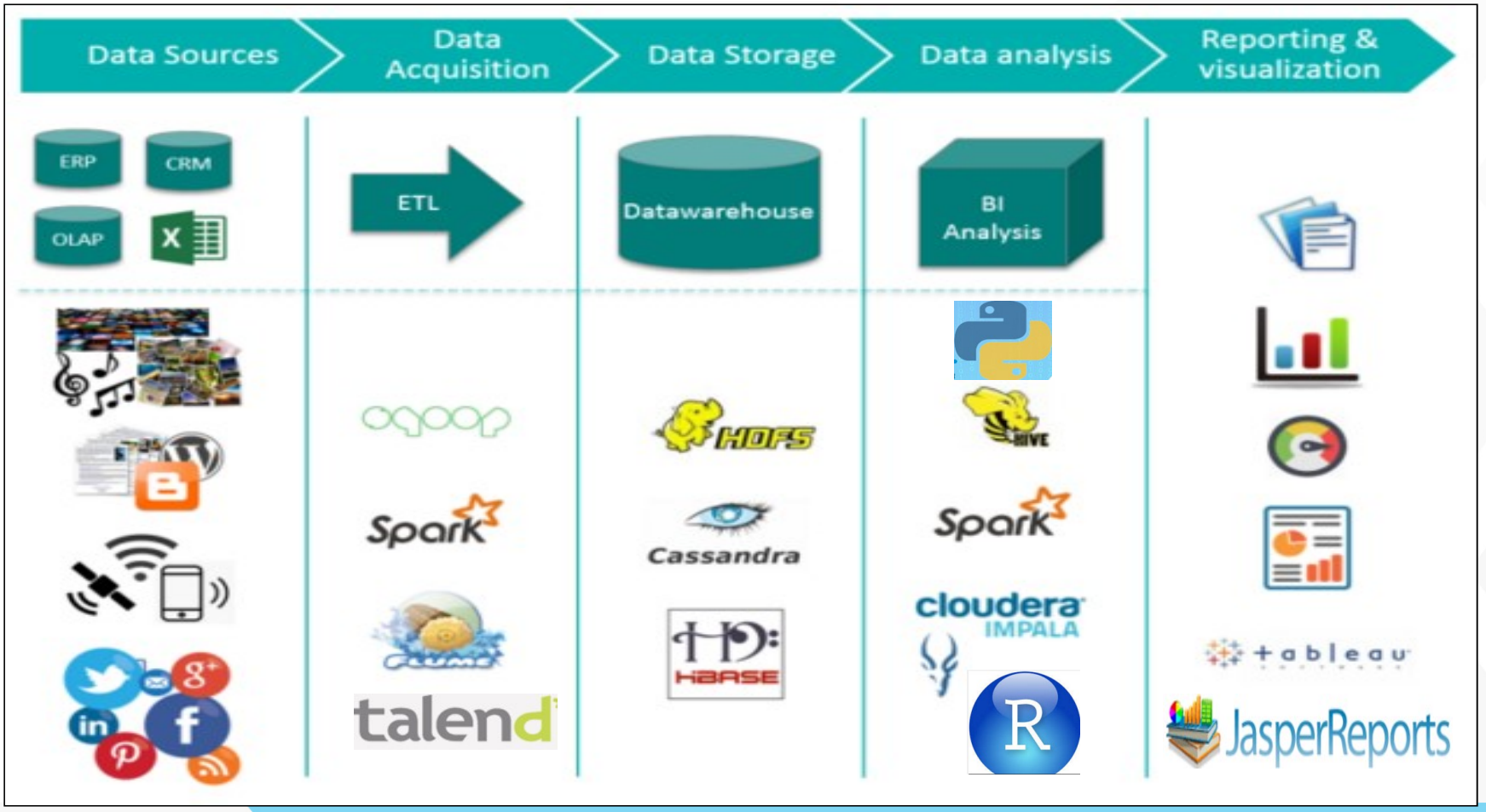
Processing

Unstructured & Semi structured

Structured

✓ Zookeeper is software project, providing an open source distributed configuration service and synchronization service and naming registry for large distributed system

Typical Big Data Project – Role of Hadoop Ecosystem



Building Blocks for Big Data Project

- Working knowledge on Hadoop & Hadoop Ecosystem
  - Be comfortable with basic Linux commands
  - Dataware housing Knowledge and SQL commands
  - Programming concepts like Java, Python, R, Pearl etc.
- Understanding data structure & Business objective
- Data visualization tools like Tableau, Qlickview, Kibana etc.
- Be comfortable with analytics tools like R, Python, Spark, SAS etc.
- Be comfortable with statistics (exploratory) and machine learning algorithms

Hadoop Symbols and Reasons Behind

Hadoop Components and Sub projects	Comparing with	Reason for the specific animal	Image
1.Hadoop distributed file system (HDFS)	Elephant	Memory of an elephant is compared with huge data storage of HDFS	
2.MapReduce	Mammoth	Mammoth means enormous, huge, massive and immense. A Mammoth's task is compared with a programming model for performing the tasks with huge volumes of data	
3.Hive	Honey Bee	Storage area of the honey in wax honeycombs inside the beehive is compared with data warehouse for storing the data in the format of table.	
4.Hbase	Horse	Running Speed of the horse indicates the real time read/write access from HBase	
5.Pig	Pig	Pigs are omnivores animals which means they can consume both plants and animals. This PIG consumes any type of data whether structured or unstructured or any other machine data and helps processing the same.	
6.Hue	Elephant foot prints	Elephant foot print is compared with "TREAD ON" Hadoop and explore it.	
7.Beeswax	Beeswax	Data storage happens in a structured way as honey stored in Beeswax	

Opportunity and Market Outlook

SECTOR	EXAMPLE APPLICATIONS	MAJOR DRIVER
Smart buildings	Automated monitoring of heating, ventilation and cooling	Reduced energy costs
Smart cities	Street lights that dim when roads are empty	Cost savings
Automotive	Emergency calling and accident alerts	
Leisure	Leisure vehicle and boat tracking	Safety and security
Consumer electronics	Connected satellite navigation devices to monitor traffic jams	Production innovation
Health	Remote monitoring of patients and personal health monitoring	Cheaper, home-based care
Utilities	Smart meters and energy demand response	Regulatory requirement
Transportation and logistics	Fleet optimization and supply-chain tracking and tracing	Cost savings
Retail	Wireless payments	Retail innovation
Manufacturing	Predictive maintenance through improved system monitoring	Reduced maintenance costs
Construction	Monitoring usage of equipment to improve efficiency and cut fuel usage	Cost savings
Agriculture and extraction	Remote monitoring of farm or mining operations and equipment	Proactive maintenance
Emergency services and national security	Disaster response and critical infrastructure protection	Faster response times

Who is using Hadoop?

Advanced analytics applications

	Single View Improve acquisition and retention	Predictive Analytics Identify your next best action	Data Discovery Uncover new findings
Financial Services	New Account Risk Screens	Trading Risk	Insurance Underwriting
	Improved Customer Service	Insurance Underwriting	Aggregate Banking Data as a Service
	Cross-sell & Upsell of Financial Products	Risk Analysis for Usage-Based Car Insurance	Identify Claims Errors for Reimbursement
Telecom	Unified Household View of the Customer	Searchable Data for NPTB Recommendations	Protect Customer Data from Employee Misuse
	Analyze Call Center Contacts Records	Network Infrastructure Capacity Planning	Call Detail Records (CDR) Analysis
	Inferred Demographics for Improved Targeting	Proactive Maintenance on Transmission Equipment	Tiered Service for High-Value Customers
Retail	360° View of the Customer	Supply Chain Optimization	Website Optimization for Path to Purchase
	Localized, Personalized Promotions	A/B Testing for Online Advertisements	Data-Driven Pricing, improved loyalty programs
	Customer Segmentation	Personalized, Real-time Offers	In-Store Shopper Behavior
Manufacturing	Supply Chain and Logistics	Optimize Warehouse Inventory Levels	Product Insight from Electronic Usage Data
	Assembly Line Quality Assurance	Proactive Equipment Maintenance	Crowdsource Quality Assurance
	Single View of a Product Throughout Lifecycle	Connected Car Data for Ongoing Innovation	Improve Manufacturing Yields
Healthcare	Electronic Medical Records	Monitor Patient Vitals in Real-Time	Use Genomic Data in Medical Trials
	Improving Lifelong Care for Epilepsy	Rapid Stroke Detection and Intervention	Monitor Medical Supply Chain to Reduce Waste
	Reduce Patient Re-Admittance Rates	Video Analysis for Surgical Decision Support	Healthcare Analytics as a Service
Oil & Gas	Unify Exploration & Production Data	Monitor Rig Safety in Real-Time	Geographic exploration
	DCA to Slow Well Declines Curves	Proactive Maintenance for Oil Field Equipment	Define Operational Set Points for Wells
Government	Single View of Entity	CBM & Autonomic Logistic Analysis	Sentiment Analysis on Program Effectiveness
	Prevent Fraud, Waste and Abuse	Proactive Maintenance for Public Infrastructure	Meet Deadlines for Government Reporting

Which companies Implemented Hadoop?

- ✓ **Its huge list**

http://wiki.apache.org/hadoop/poweredBy

Hadoop 2.7.0

Apache Hadoop-2.7.0- Components(1/2)

- The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.
- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

- The project includes these modules:
  - **Hadoop Common**: The common utilities that support the other Hadoop modules.
  - **Hadoop Distributed File System (HDFS)**: a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
  - **Hadoop YARN**: a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications.
  - **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets(programming model for large scale data processing)

Apache Hadoop-2.7.0- Components(2/2)

- There are five pillars to Hadoop that make it enterprise ready:
  1. Data Management: **Apache Hadoop YARN**, **HDFS**
  2. Data Access: **Apache Hive**, **Apache Pig**, **MapReduce**, **Apache Spark**, **Apache Storm**, **Apache Hbase**, **Apache Tez**, **Apache Kafka**, **Apache Hcatalog**, **Apache Slider**, **Apache Solr**, **Apache Mahout**, **Apache Accumulo**
  3. Data Governance and Integration: **Apache Falcon**, **Apache Flume**, **Apache Sqoop**
  4. Security: **Apache Knox**, **Apache Ranger**
  5. Operations: **Apache Ambari**, **Apache Oozie**, **Apache ZooKeeper**

Hadoop Ecosystem– Providers(Vendors)

Commercial Vendors

- Cloudera
- Hortonworks
- IBM Infosphere Biginsights
- MapR Technologies
- Think Big Analytics
- Amazon Web Services (Cloud based)
- Microsoft Azure (Cloud based)

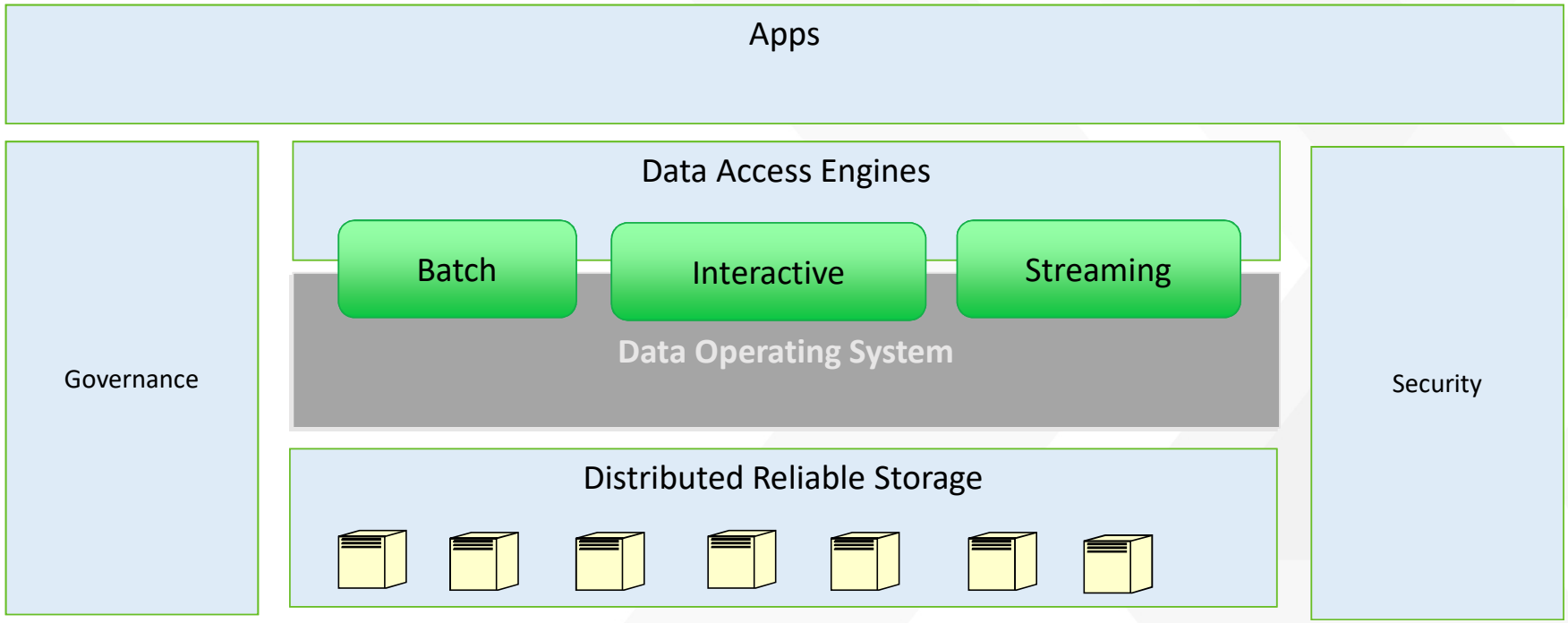
Open Source Vendors

- Apache
- Apache Bigtop
- Cascading
- Cloudspace
- Datameer
- Data Mine Lab
- Data Salt
- Data Stax
- Data Torrent
- Debian
- Emblocsoft
- Hstreaming
- Impetus
- Pentaho
- Talend
- Jaspersoft
- Karmasphere
- Apache Mahoot
- Nutch
- NGData
- Pervasive Software
- Pivotal
- Sematext International
- Syncsort
- Tresata
- Wandisco
- Etc..

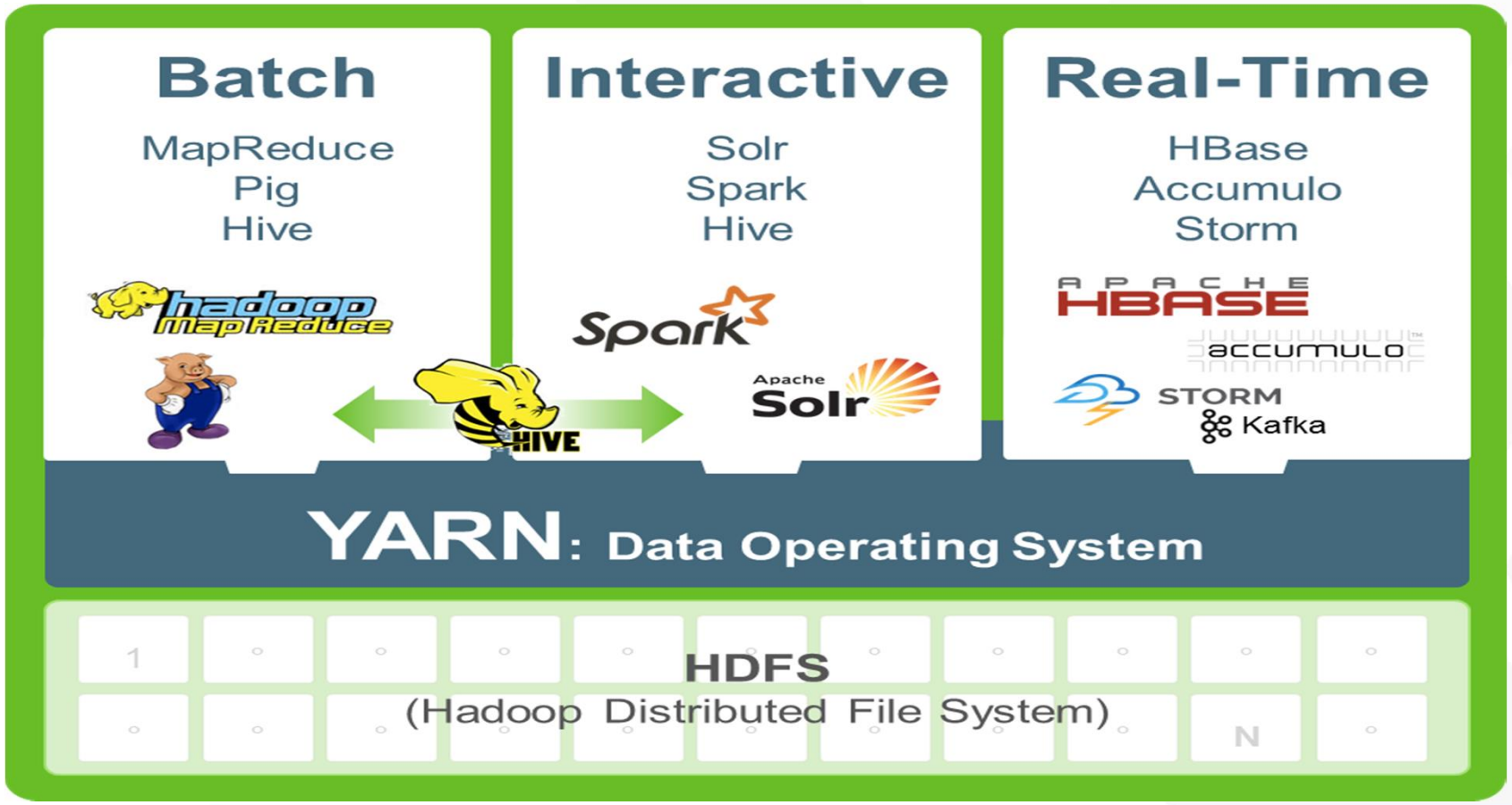
http://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support

Hadoop - Architecture

Hadoop Architecture



Hadoop Architecture



Hadoop Architecture

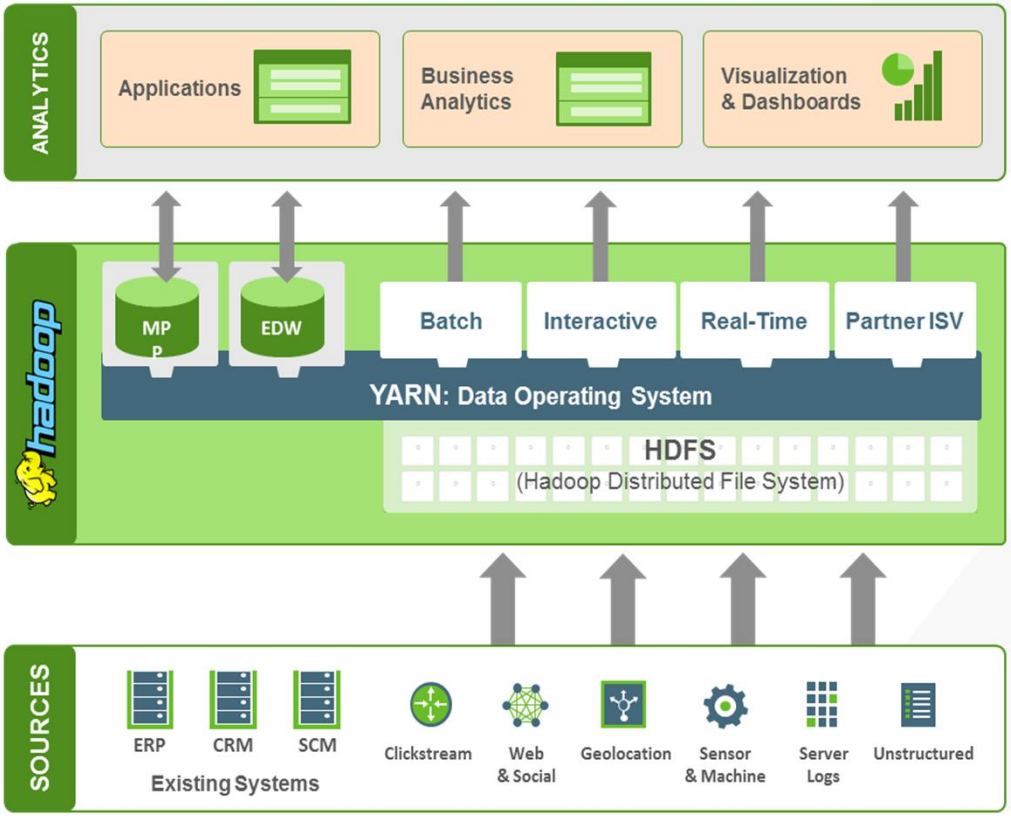
Batch Processing
Batch processing is the processing of previously collected data in the form of batches i.e., processing large amounts of accumulated data at one time. Hadoop handles the large amounts of data by storing them in a cluster of nodes connected to each other.

Real-time Processing
Real time processing is the data processing that takes place, instantaneously upon data entry or receipt of a command. Real-time processing must execute within strict constraints on response time.

Near Real-time Processing
Here there is a slight time delay between the collection of data and processing of data.

HDFS+MapReduce- Batch processing
HDFS+Storm – Real-time processing
HDFS+Spark – Near real-time processing

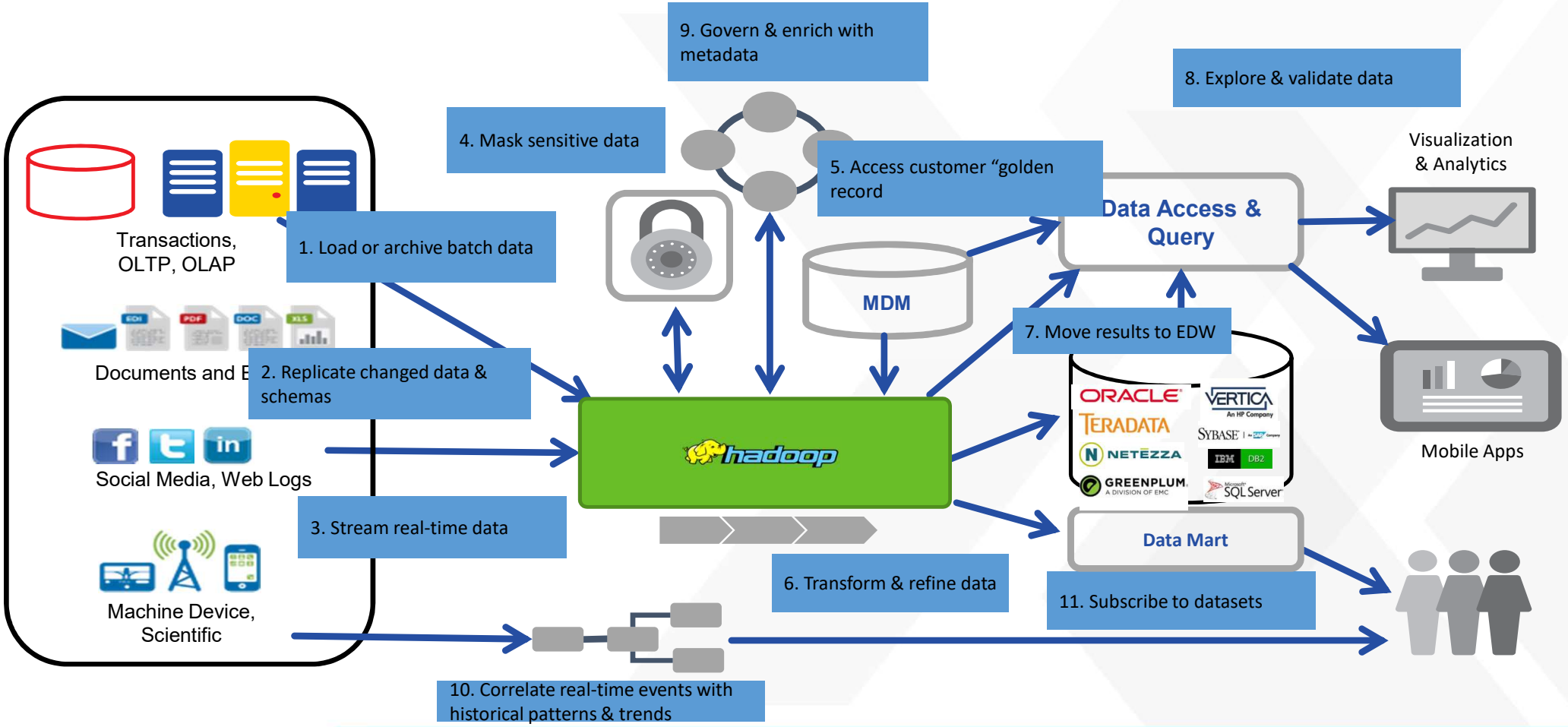
Modern Data Architecture emerges to unify data & processing



Modern Data Architecture

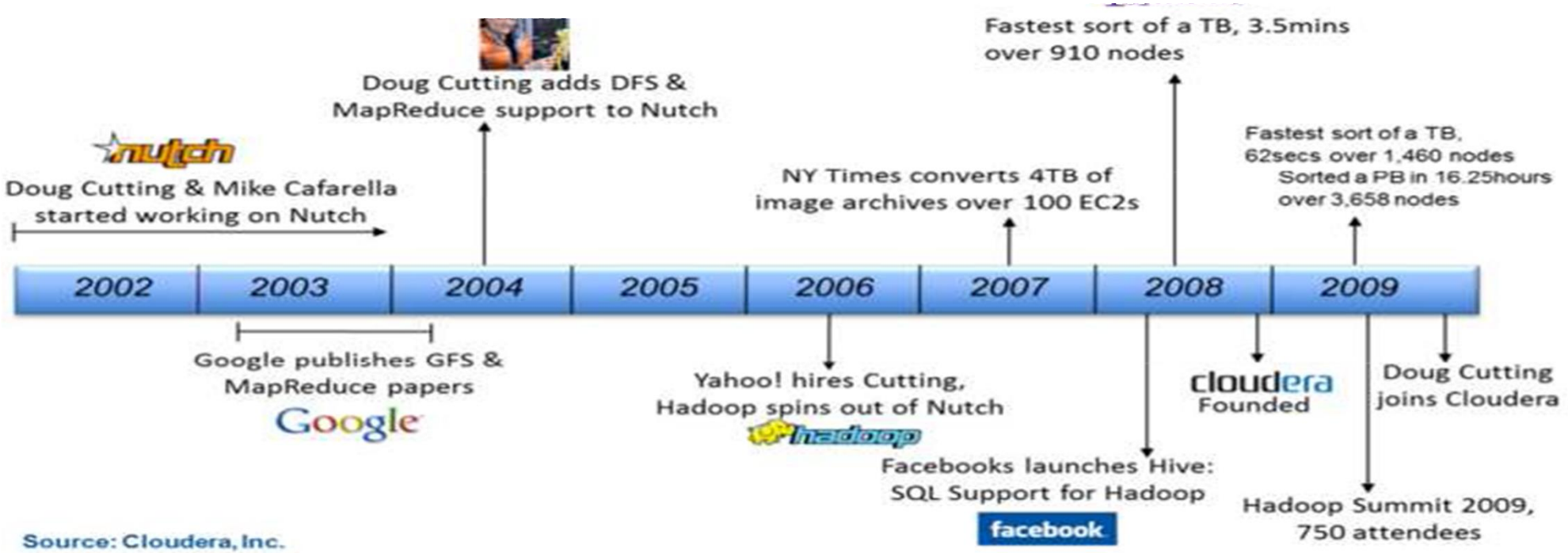
- Enable applications to have access to all your enterprise data through an efficient centralized platform
- Supported with a centralized approach governance, security and operations
- Versatile to handle any applications and datasets no matter the size or type

Big Data ETL Life Cycle



Appendix

Hadoop Timeline



Hadoop: brief history

- Hadoop was created by Doug Cutting and Mike Cafarella
- Name behind his son's toy elephant
- 1 billion page indexing would cost around half a million dollors in hardware and $30000 monthly running cost
- It was originally developed to support distribution for the Nutch search engine project
- GFS(Google File System) => NDFS (Nutch Distributed File System)
- Nutch Joined Yahoo
- Hadoop dedicated to open source Apache Foundation
- 2008, Hadoop broke record and become the fastest system to sort terabyte of data
- Hadoop sorted 1 terabyte of data in 209 second by running 910 node cluster beating previous record of 297 second
- In November same year Google reported that its MapReduce implementation sorted 1 terabyte in 68 second

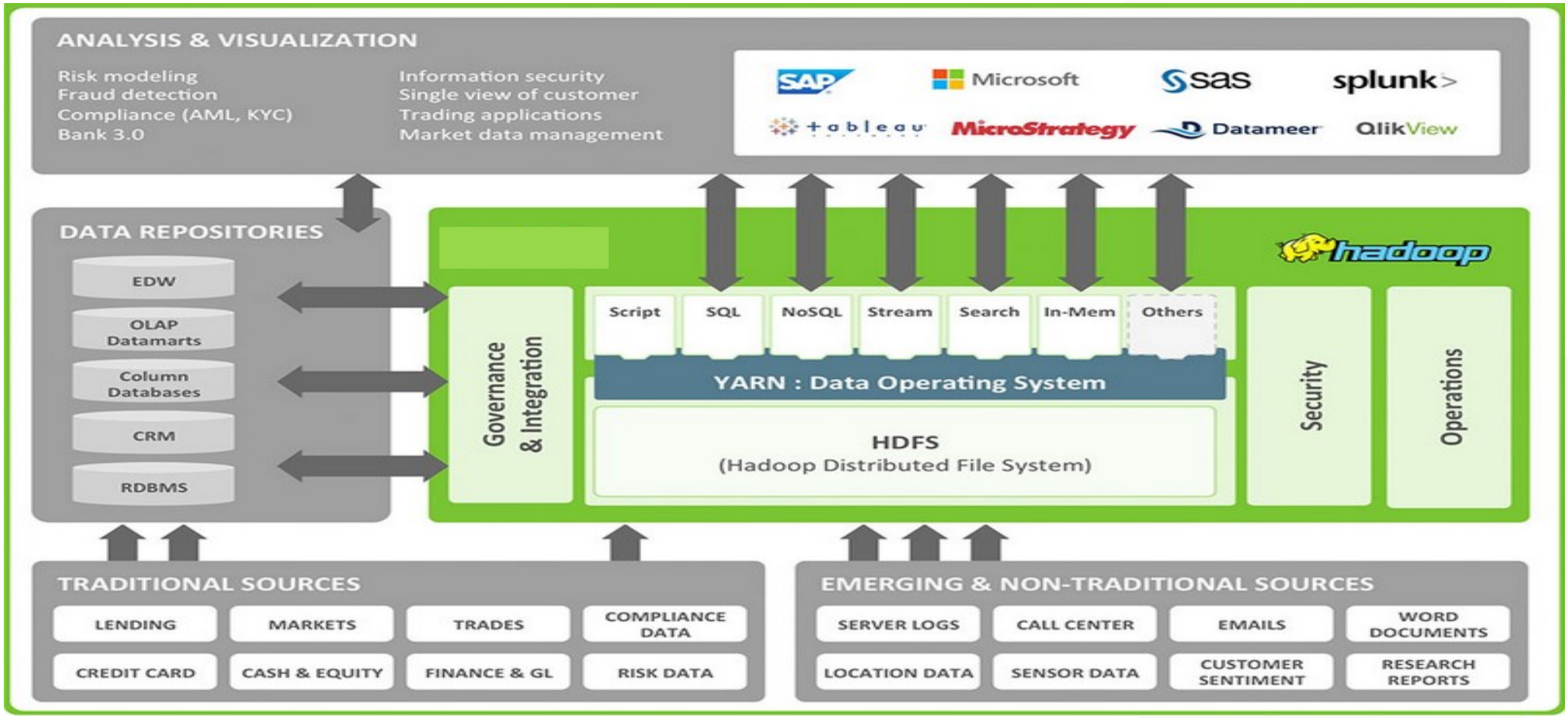
Hadoop: brief history cont…

- 2004, Initial versions of what is now Hadoop Distributed Filesystem and MapReduce implemented by Doug Cutting and Mike Caferella
- Dec. 2005, Nutch ported to the new framework. Hadoop runs reliably on 20 nodes
- January 2006 Doug Cutting joins Yahoo!
- February 2006 apache Hadoop project officially started to support the stand-alone development of MapReduce and HDFS
- February 2006 Adoption of Hadoop by Yahoo Grid team
- April 2006 Sort benchmark(10gb/node run on 188 nodes in 47.9 hours
- May 2006 Yahoo setup a Hadoop research cluster 300 nodes
- May 2006 Sort benchmark run on 500 nodes in 42 hours(better hardware than April benchmark)
- Oct. 2006 Research cluster reaches 600 nodes

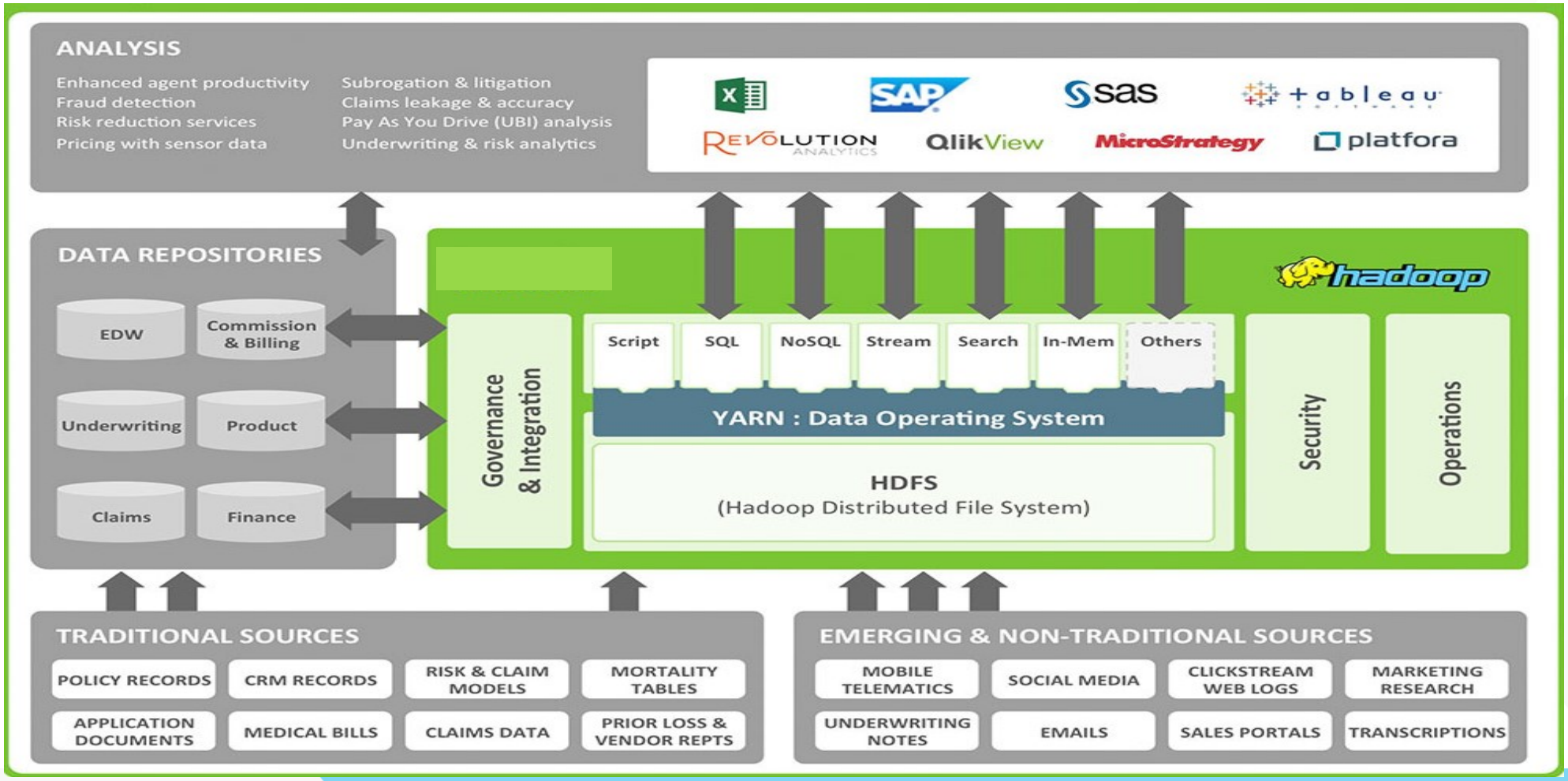
Hadoop: brief history cont…

- Dec. 2006 sort benchmark run on 20 nodes in 1.8 hours, 100 nodes in 3.3 hours, 500 nodes in 5.2 hours, 900 nodes in 7.8 hours
- Jan 2007 Research cluster reaches 900 nodes
- April 2007 Research cluster 2 cluster of 1000 nodes
- April 2008 Won the 1 terabyte sort benchmark in 2009 sec on 900 nodes
- Oct.2008 Loading 10 terabytes of data per day on to research clusters
- March 2009 17 cluster with a total of 24000 nodes
- April 2009 won the minute sort by sorting 500 gb in 59 seconds(on 1400 nodes) and 100 terabytes sort in 173 minute(on 3400 nodes)

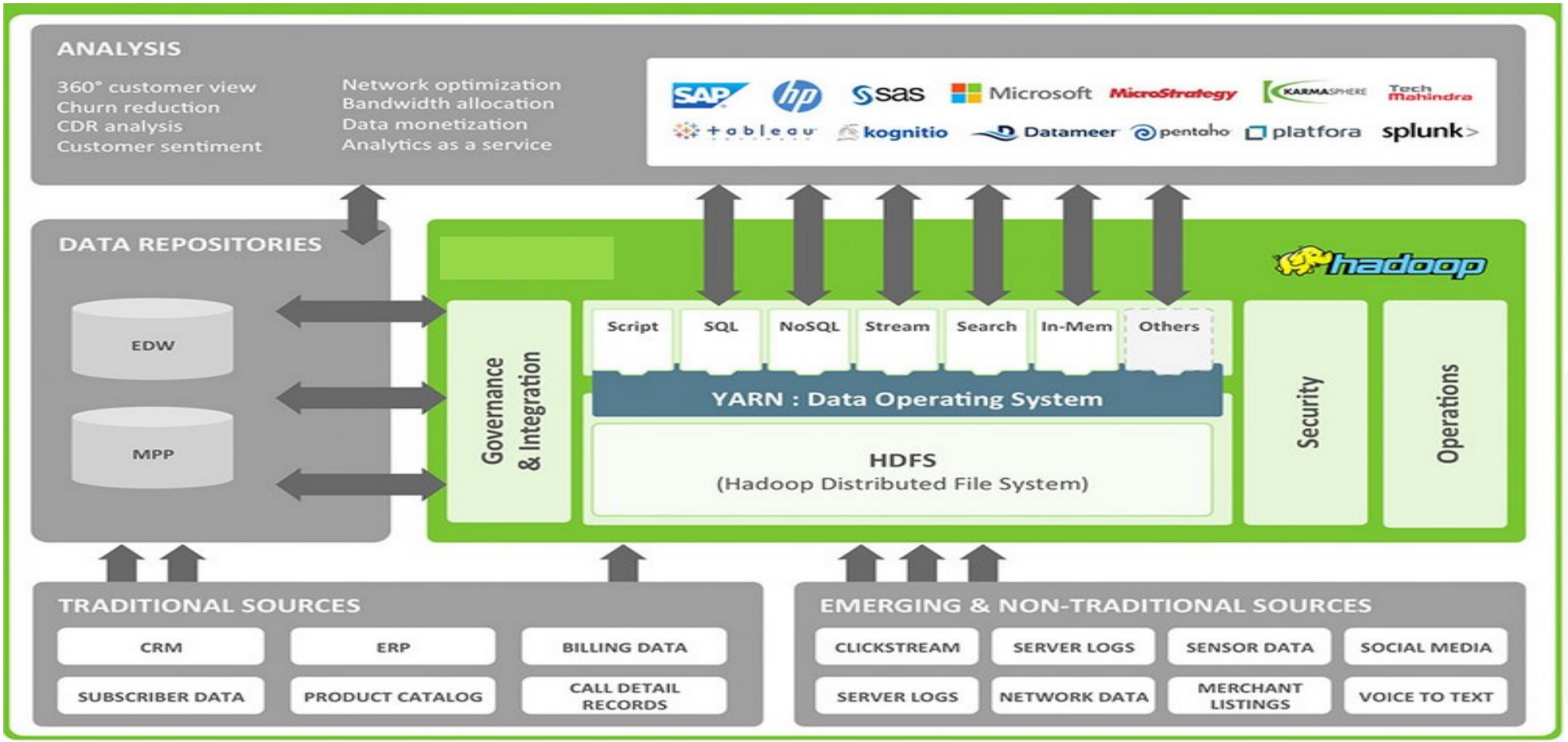
Hadoop for financial services



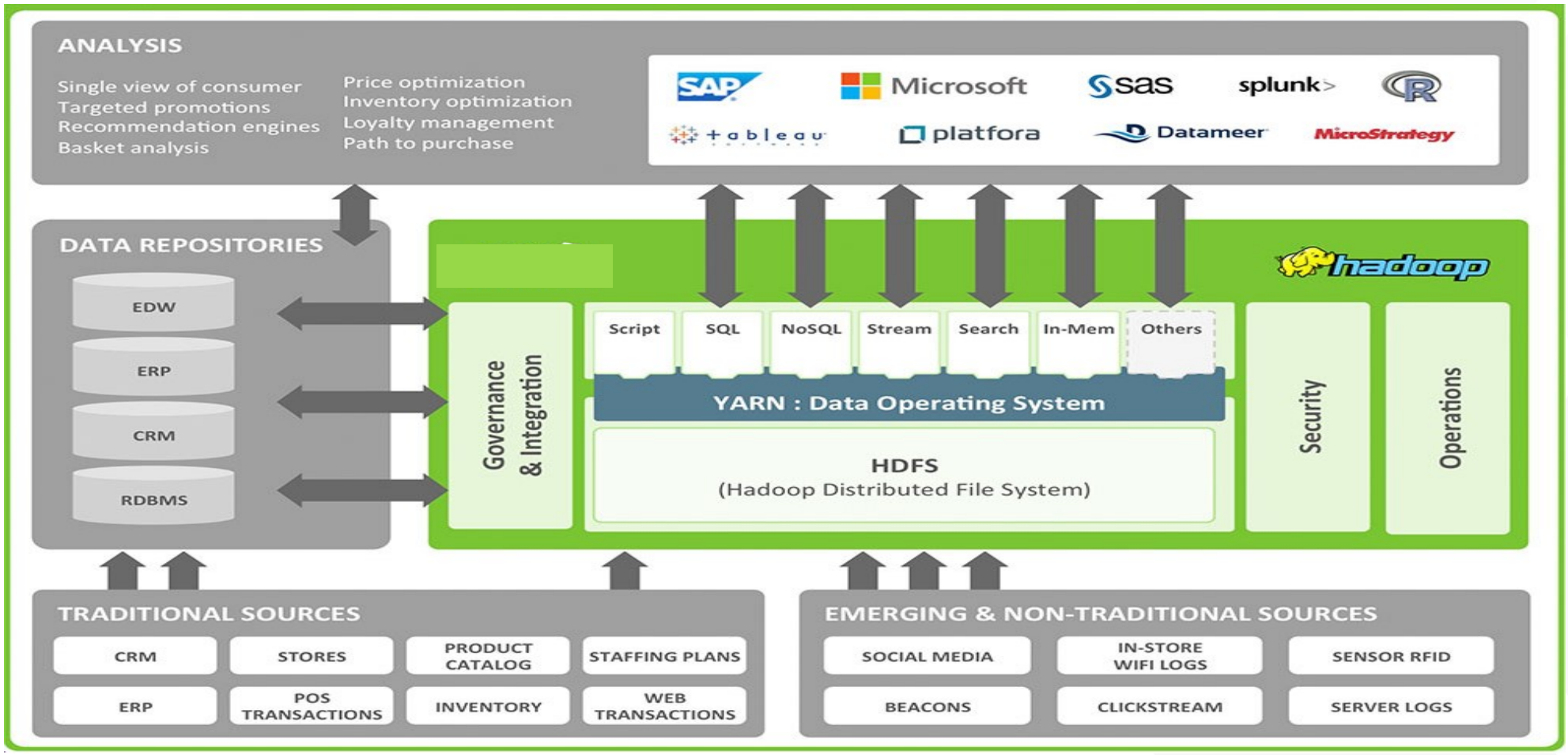
Hadoop for Insurance



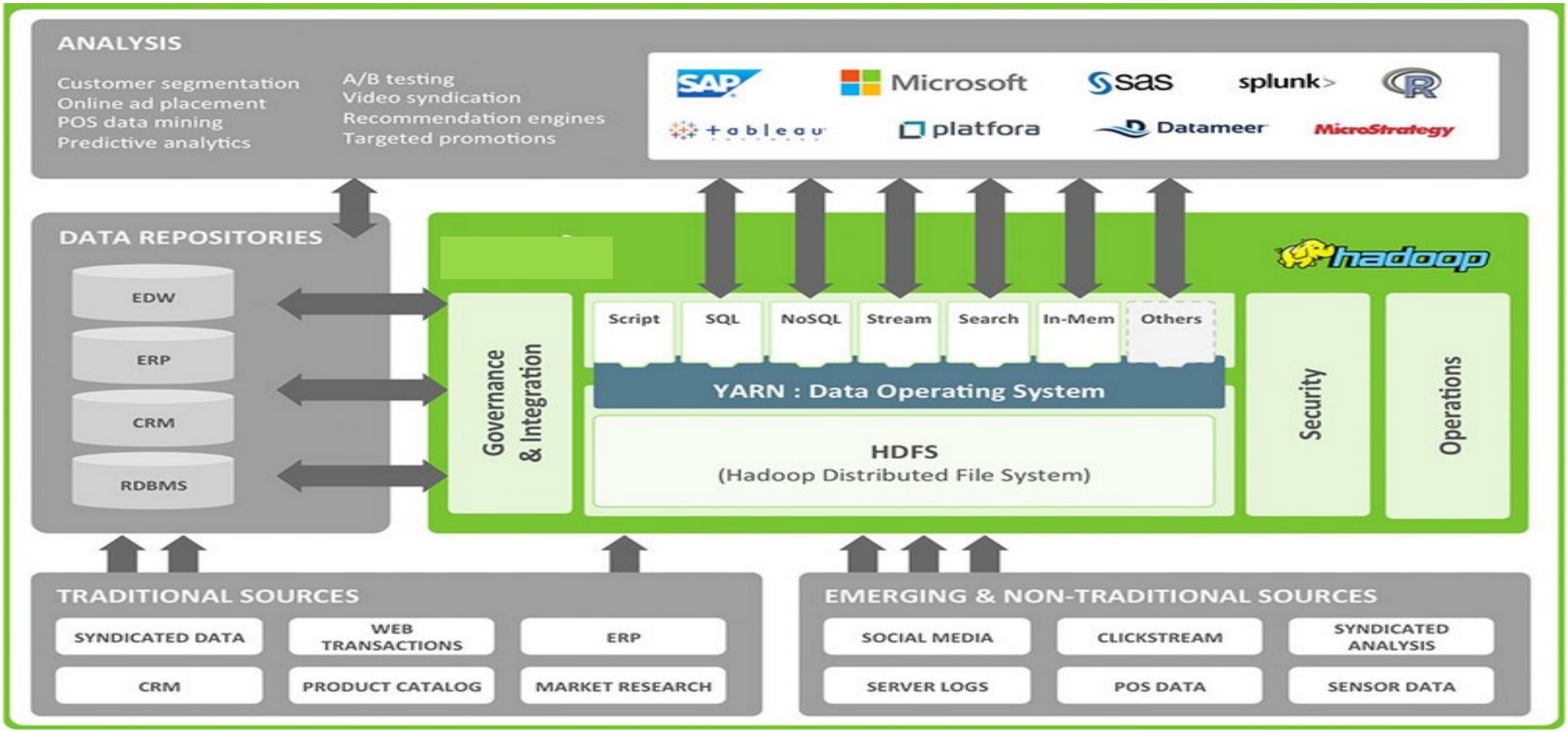
Hadoop for Telecommunications



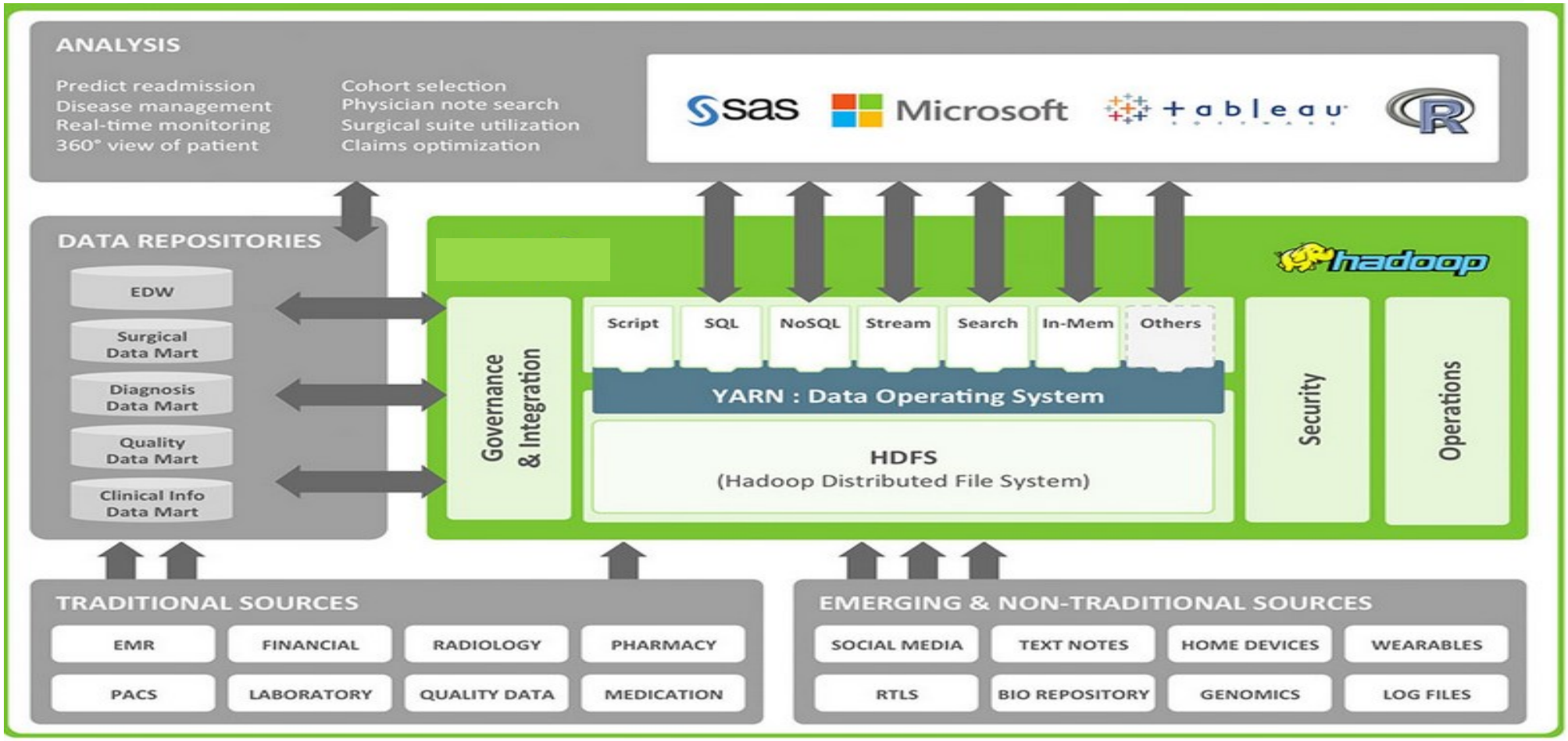
Hadoop for Retailors



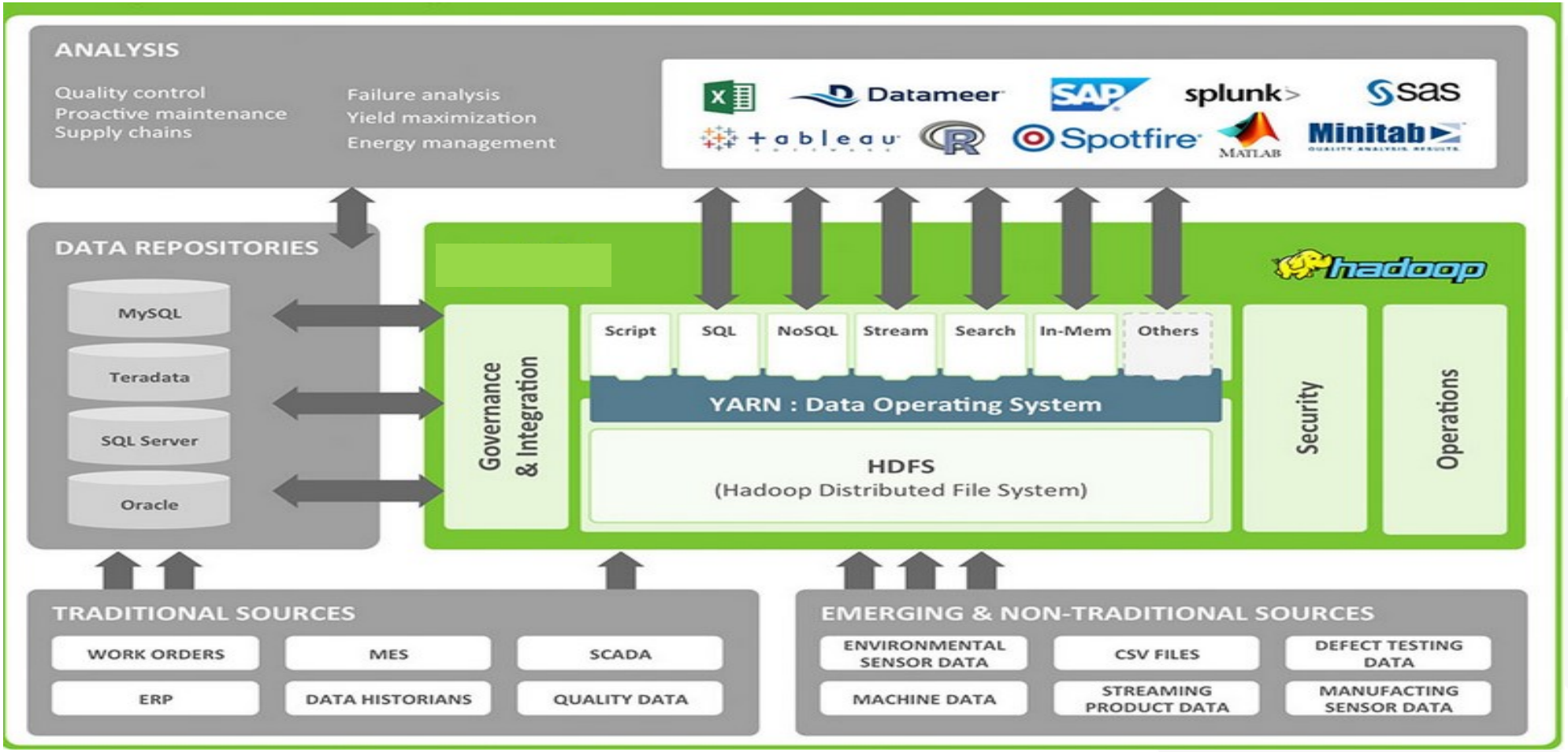
Hadoop for Advertising



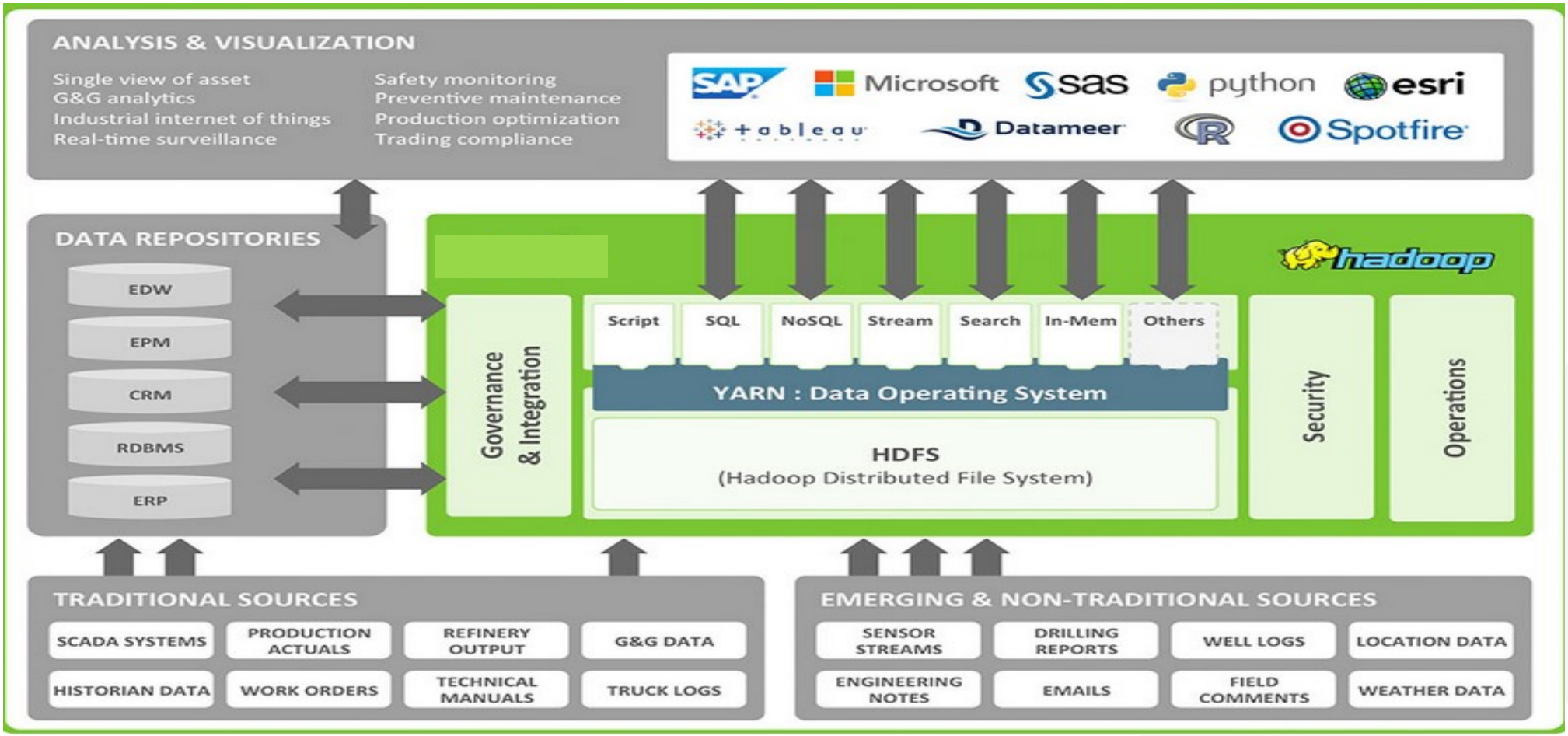
Hadoop for Healthcare



Hadoop for Manufacturing



Hadoop for Oil & Gas



Data Management

- **Data Management:**
  - Store and process vast quantities of data in a storage layer that scales linearly. Hadoop Distributed File System (HDFS) is the core technology for the efficient scale out storage layer, and is designed to run across low-cost commodity hardware.
  - Apache Hadoop YARN is the pre-requisite for Enterprise Hadoop as it provides the resource management and pluggable architecture for enabling a wide variety of data access methods to operate on data stored in Hadoop with predictable performance and service levels.

- Apache Hadoop YARN– Part of the core Hadoop project, YARN is a next-generation framework for Hadoop data processing extending MapReduce capabilities by supporting non-MapReduce workloads associated with other programming models.

- HDFS– Hadoop Distributed File System (HDFS) is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers

Data Access

Data Access–

- Interact with your data in a wide variety of ways – from batch to real-time.
- Apache Hive is the most widely adopted data access technology, though there are many specialized engines.

Apache Hive– Built on the MapReduce framework, Hive is a data warehouse that enables easy data summarization and ad-hoc queries via an SQL-like interface for large datasets stored in HDFS.

Apache Pig– A platform for processing and analyzing large data sets. Pig consists of a high-level language (Pig Latin) for expressing data analysis programs paired with the MapReduce framework for processing these programs.

MapReduce– MapReduce is a framework for writing applications that process large amounts of structured and unstructured data in parallel across a cluster of thousands of machines, in a reliable and fault-tolerant manner.

Apache Spark– Spark is ideal for in-memory data processing. It allows data scientists to implement fast, iterative algorithms for advanced analytics such as clustering and classification of datasets.

Data Access

Apache Storm– Storm is a distributed real-time computation system for processing fast, large streams of data adding reliable real-time data processing capabilities to Apache Hadoop® 2.x

Apache HBase– A column-oriented NoSQL data storage system that provides random real-time read/write access to big data for user applications.

Apache Tez– Tez generalizes the MapReduce paradigm to a more powerful framework for executing a complex DAG (directed acyclic graph) of tasks for near real-time big data processing.

Apache Kafka– Kafka is a fast and scalable publish-subscribe messaging system that is often used in place of traditional message brokers because of its higher throughput, replication, and fault tolerance.

Apache HCatalog– A table and metadata management service that provides a centralized way for data processing systems to understand the structure and location of the data stored within Apache Hadoop.

Data Access

Apache Slider– A framework for deployment of long-running data access applications in Hadoop. Slider leverages YARN's resource management capabilities to deploy those applications, to manage their lifecycles and scale them up or down.

Apache Solr– Solr is the open source platform for searches of data stored in Hadoop. Solr enables powerful full-text search and near real-time indexing on many of the world's largest Internet sites.

Apache Mahout– Mahout provides scalable machine learning algorithms for Hadoop which aids with data science for clustering, classification and batch based collaborative filtering.

Apache Accumulo– Accumulo is a high performance data storage and retrieval system with cell-level access control. It is a scalable implementation of Google's Big Table design that works on top of Apache Hadoop and Apache ZooKeeper.

Data governance & Integration

Data Governance and Integration:

- Quickly and easily load data, and manage according to policy. Apache Falcon provides policy-based workflows for data governance, while Apache Flume and Sqoop enable easy data ingestion, as do the NFS and WebHDFS interfaces to HDFS.

- Apache Falcon– Falcon is a data management framework for simplifying data lifecycle management and processing pipelines on Apache Hadoop®. It enables users to orchestrate data motion, pipeline processing,disaster recovery, and data retention workflows.

- Apache Flume– Flume allows you to efficiently aggregate and move large amounts of log data from many different sources to Hadoop.

- Apache Sqoop– Sqoop is a tool that speeds and eases movement of data in and out of Hadoop. It provides a reliable parallel load for various, popular enterprise data sources

Security

Security:

- Address requirements of Authentication, Authorization, Accounting and Data Protection. Security is provided at every layer of the Hadoop stack from HDFS and YARN to Hive and the other Data Access components on up through the entire perimeter of the cluster via Apache Knox.

- Apache Knox– The Knox Gateway ("Knox") provides a single point of authentication and access for Apache Hadoop services in a cluster. The goal of the project is to simplify Hadoop security for users who access the cluster data and execute jobs, and for operators who control access to the cluster.

- Apache Ranger– Apache Ranger delivers a comprehensive approach to security for a Hadoop cluster. It provides central security policy administration across the core enterprise security requirements of authorization, accounting and data protection.

Operations

Operations– Provision, manage, monitor and operate Hadoop clusters at scale.

Apache Ambari– An open source installation lifecycle management, administration and monitoring system for Apache Hadoop clusters.

Apache Oozie– Oozie Java Web application used to schedule Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work.

Apache ZooKeeper– A highly available system for coordinating distributed processes. Distributed applications use ZooKeeper to store and mediate updates to important configuration information.

Contact Us

Visit us on: http://www.analytixlabs.in/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: (+91) 9910509849

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/