MATES ED2MIT

Education and Training for Data Driven Maritime Industry

Tutorial A02

Big Data Platforms for Data Analytics

Yuri Demchenko MATES Project
University of Amsterdam

Maritime Alliance for fostering the European Blue economy through a Marine Technology Skilling Strategy

Co-funded by the Erasmus+ Programme of the European Union

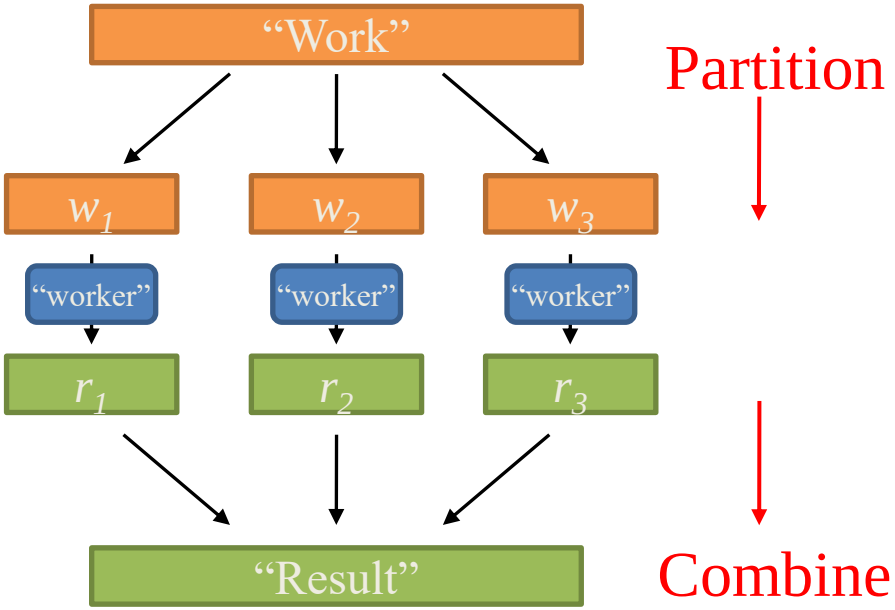
Outline

- Big Data Algorithms
  - MapReduce, Pregel
- Hadoop platform for Big Data storage and processing
- Cloud based storage for Big Data, Data Lakes
- Big Data Platforms and Providers
  - Amazon Web Services (AWS) Big Data services
  - Google Cloud Platform (GCP) Big Data services
  - Microsoft Azure Analytics Platform and HDInsight
- Demo AWS EMR deployment

- Discussion: Big Data solutions for your company

Co-funded by the
Erasmus+ Programme
of the European Union

Algorithms for Big Data Processing



- **Parallelisation and scalable computing**
- **MapReduce** Computation Model
  - Map-Reduce centric thinking
- MapReduce and **Hadoop** ecosystem
- **Pregel** algorithm and graph processing
  - Node centric thinking

Tasks typical for Big Data Analytics

- Search and ranking – for webpages, services
- Classification
- Regression
- Similarity
- Graph analytics and linking

MapReduce Programming Model

Map/Reduce is a programming model based on LISP that allows simple distribution of computing tasks across nodes. It includes two stages:

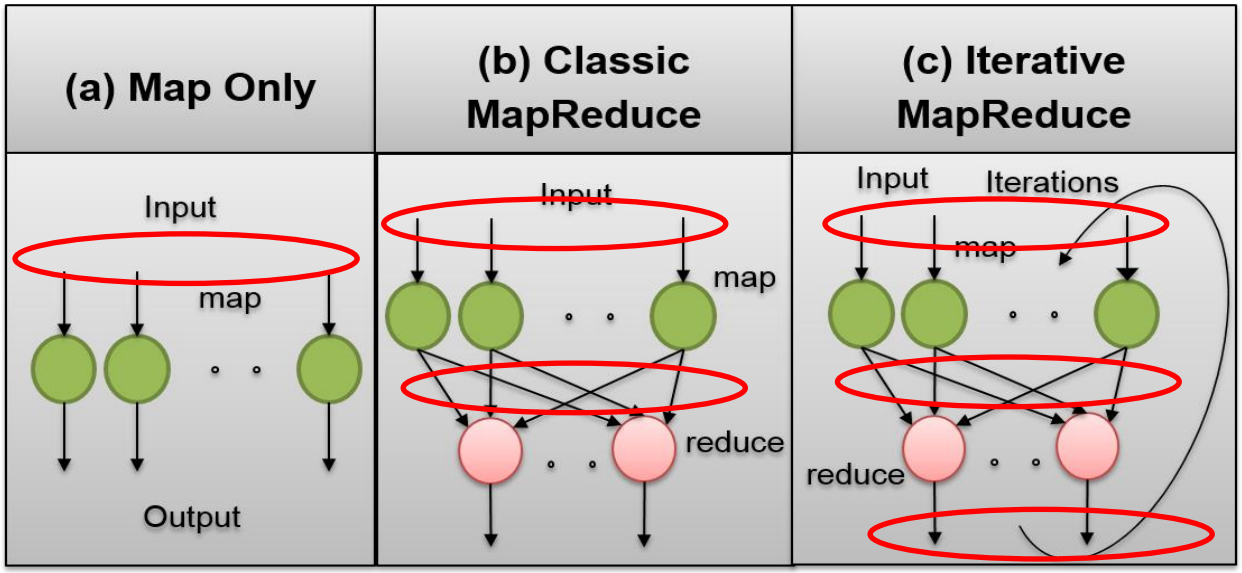
- **Map:** Perform a function on individual values in datasets to create a new list of values
- **Reduce:** Combine values from the new list to create a new value

Input and output are presented as key/value pairs. The following code fragment expresses the two operations map() and reduce() that run in parallel for two strings to execute word count:

```
map (in_key, in_value)
       list (out_key, intermediate_value)

reduce (out_key, list (intermediate_value))
       list (out_key, out_value)
```

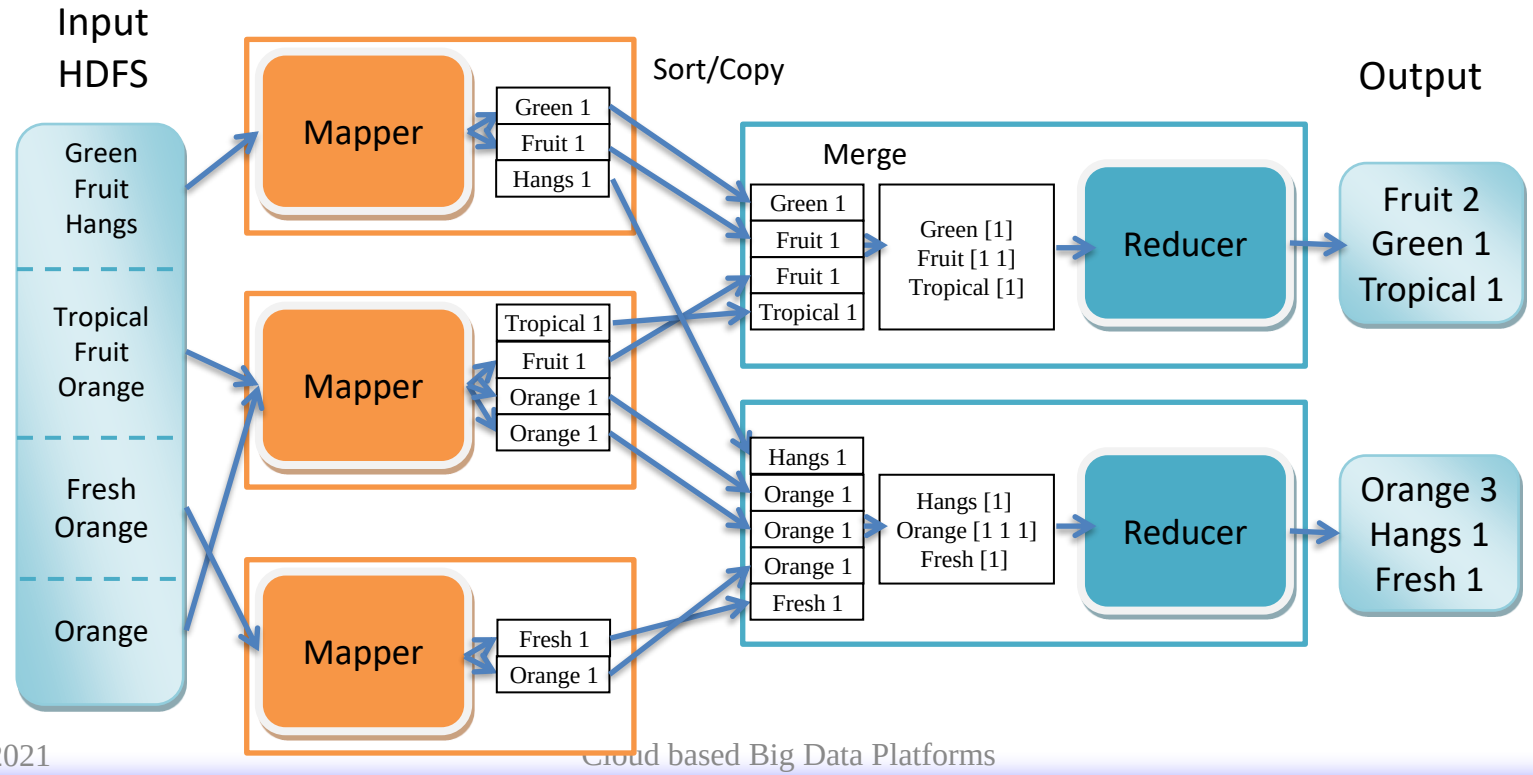
Three Forms of MapReduce



- The Map function produces an intermediate value for the intended output key, the Reduce function combines all intermediate values for a particular key.
- Data flow in different forms of MapReduce: (a) Map Only; (b) Classic MapReduce; (c) Iterative MapReduce.
- The classic MapReduce operates in a **synchronous mode**.
  - Input data are partitioned and multiple map() tasks are run in parallel.
  - After all map()s are complete, all intermediate values are combined for all unique keys by running multiple reduce() tasks in parallel.

Word Count with MapReduce

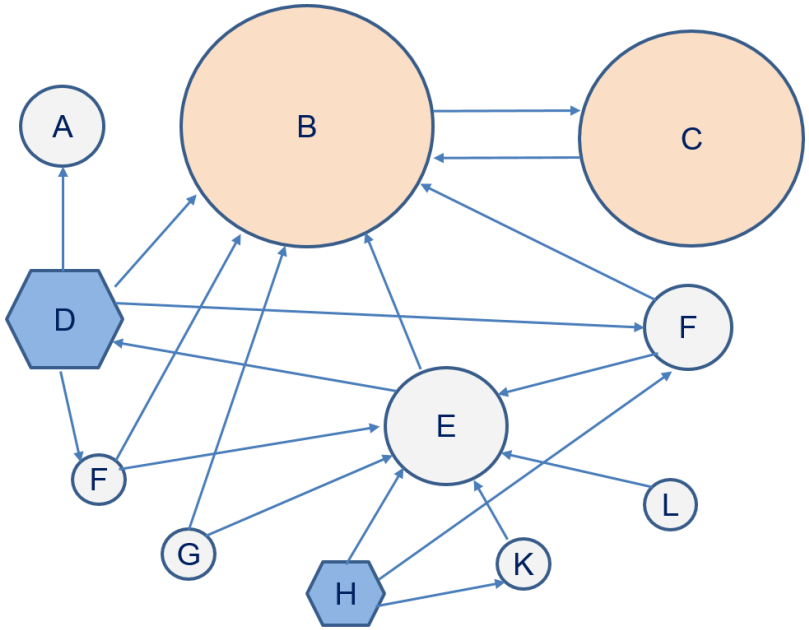
Input strings		Mapper			Reducer	
		Green	1		Fruit	2
		Fruit	1		Green	1
Green fruit hangs	➔	Hangs	1	➔	Hangs	1
Tropical fruit orange	➔			➔	Tropical	1
		Tropical	1		Orange	1
		Fruit	1			
		Orange	1			



Word Count: Try to exercise with a simple text

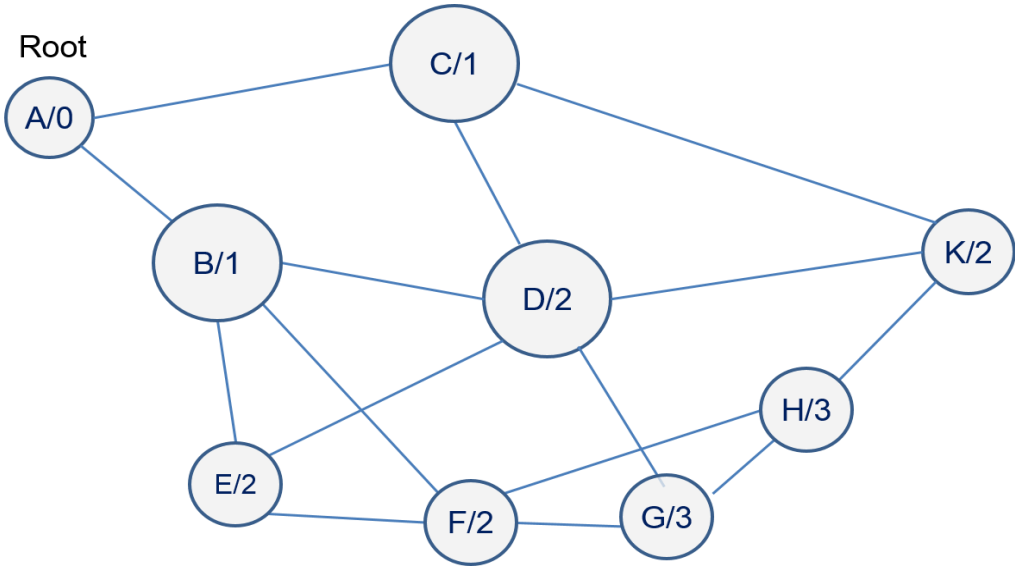
- Split text on few parts (optional) – also called partitions in Big Data
- Split all text on words <-- Map
- Order and group words
- Combine similar works and count <-- Reduce
- Store results
- Do it manually, with Excel or use python or Java program
- Practical aspects when analysing web pages
  - Mind stop words and word stemming
    - Stop words are typically removed
    - Often different form of the same stem are combined
  - Text extraction from web page in HTML form
    - Special content in HTML page: Metadata, visible/invisible text, etc

Websearch technics: Page Ranking Example



- A,B,C…L pages presented as graph nodes
- **Hub pages** (shown as hexagons) and **authority pages** (shown as circles).
- The size of the page is proportional to its rank.
- Search Engine results typically show authoritative pages at the top of search results page.
  - The highest ranked page is the first.
- It is a part of the Search Engines business to add paid advertisement on the search page but they are typically located in a special area like this is done by Google.

Pregel Algorithm: Calculation of the shortest path in the graph



- Showing shortest path to the Root
- Edges weight is 1
- Edges are unidirectional

- Used for social graph calculation

Discussion and Polling

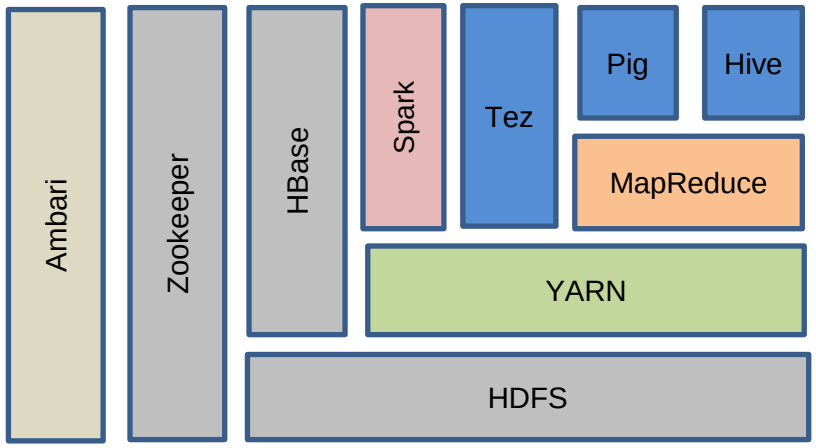
- Go to **www.menti.com**

- Use of algorithms and applications

MapReduce and Hadoop

- Java based
- Designed for scalability
  - Up to TB data: distributed, not-consistent
- And not for speed
  - Even simple data query task will take seconds

Apache Hadoop (Release 2.2+)



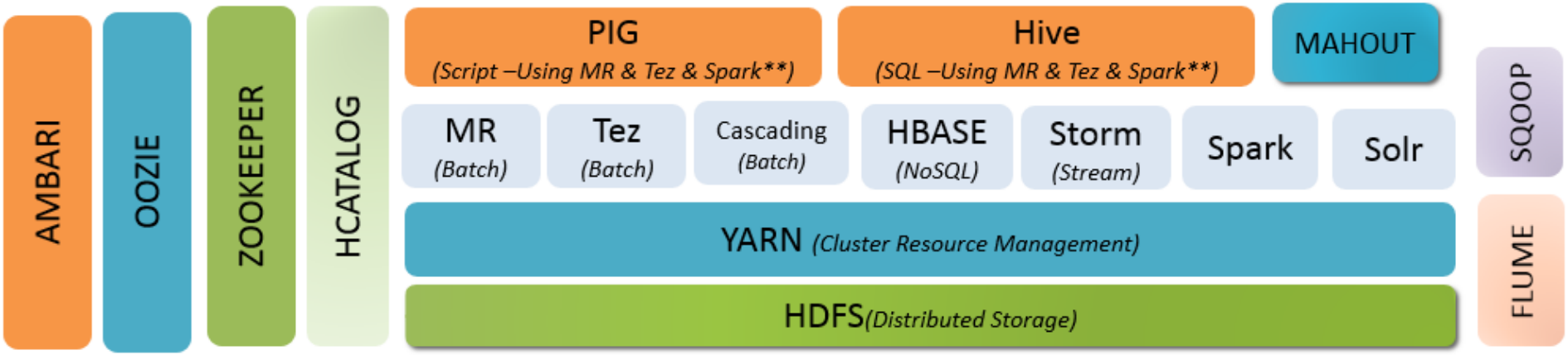
Apache Hadoop software stack includes the following main modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules and includes utilities and drivers to support different computer cluster and language platforms.
- **HDFS**: Hadoop Distributed File System optimized for large scale storage and processing of data on commodity hardware
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

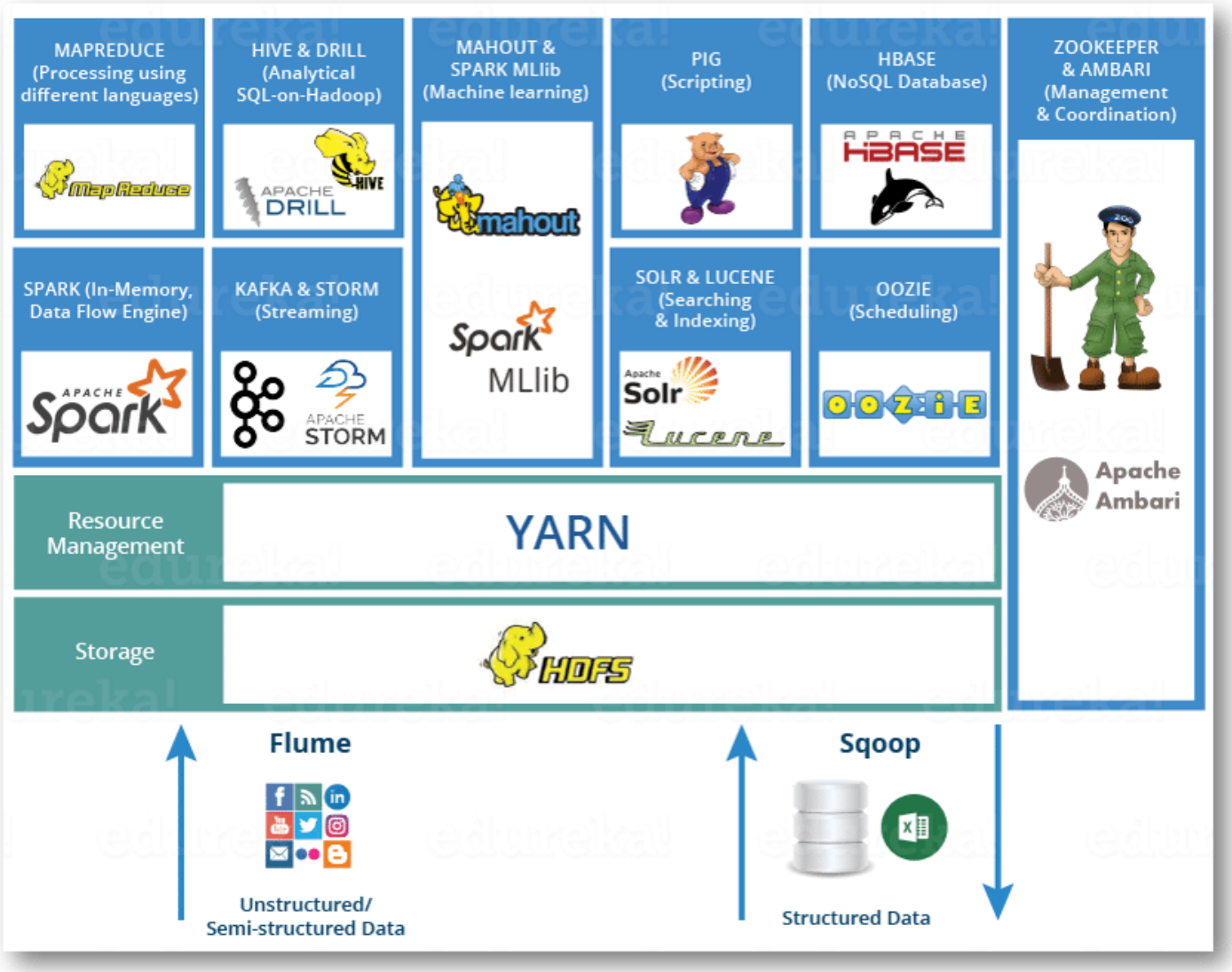
Other Hadoop-related projects at Apache include:

- **Hive:** A data warehouse system that provides data aggregation and querying.
- **Pig:** A high-level data-flow language and execution framework for parallel computation.
- **HBase:** A distributed column oriented database that supports structured data storage for large tables
- **Tez:** A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.
- **ZooKeeper:** A scalable coordination service for distributed applications.
- **Spark:** A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Ambari:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters
- **Cassandra:** A scalable multi-master database protected against hardware failure
- **Mahout:** A scalable machine learning and data mining library.
- **Avro:** A data serialization system that supports rich data structures

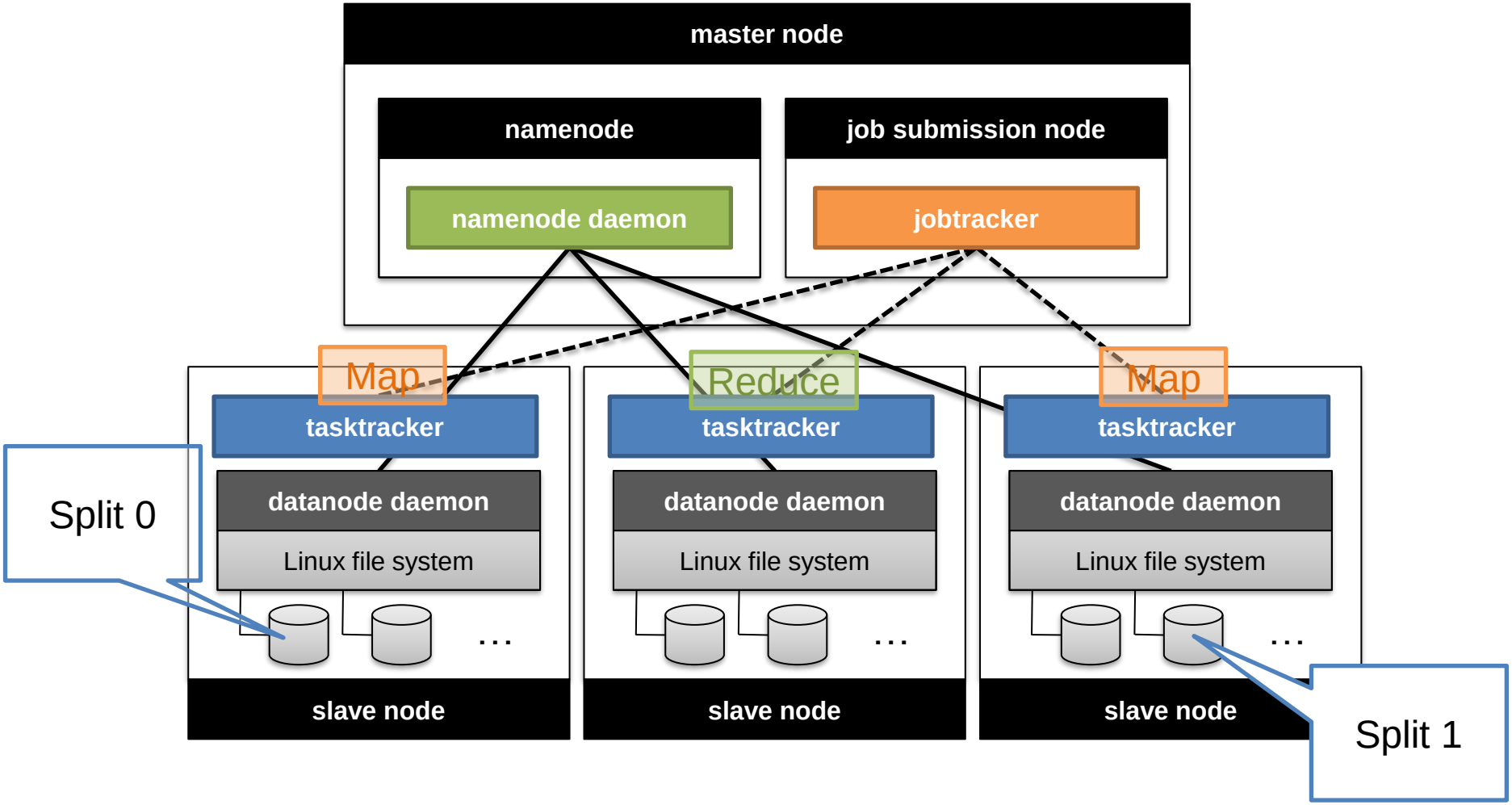
Hadoop Ecosystem – Layered view



Zoo style Hadoop Ecosystem



Hadoop Cluster Architecture – “Physical” view



Hadoop Cluster operation on MapReduce

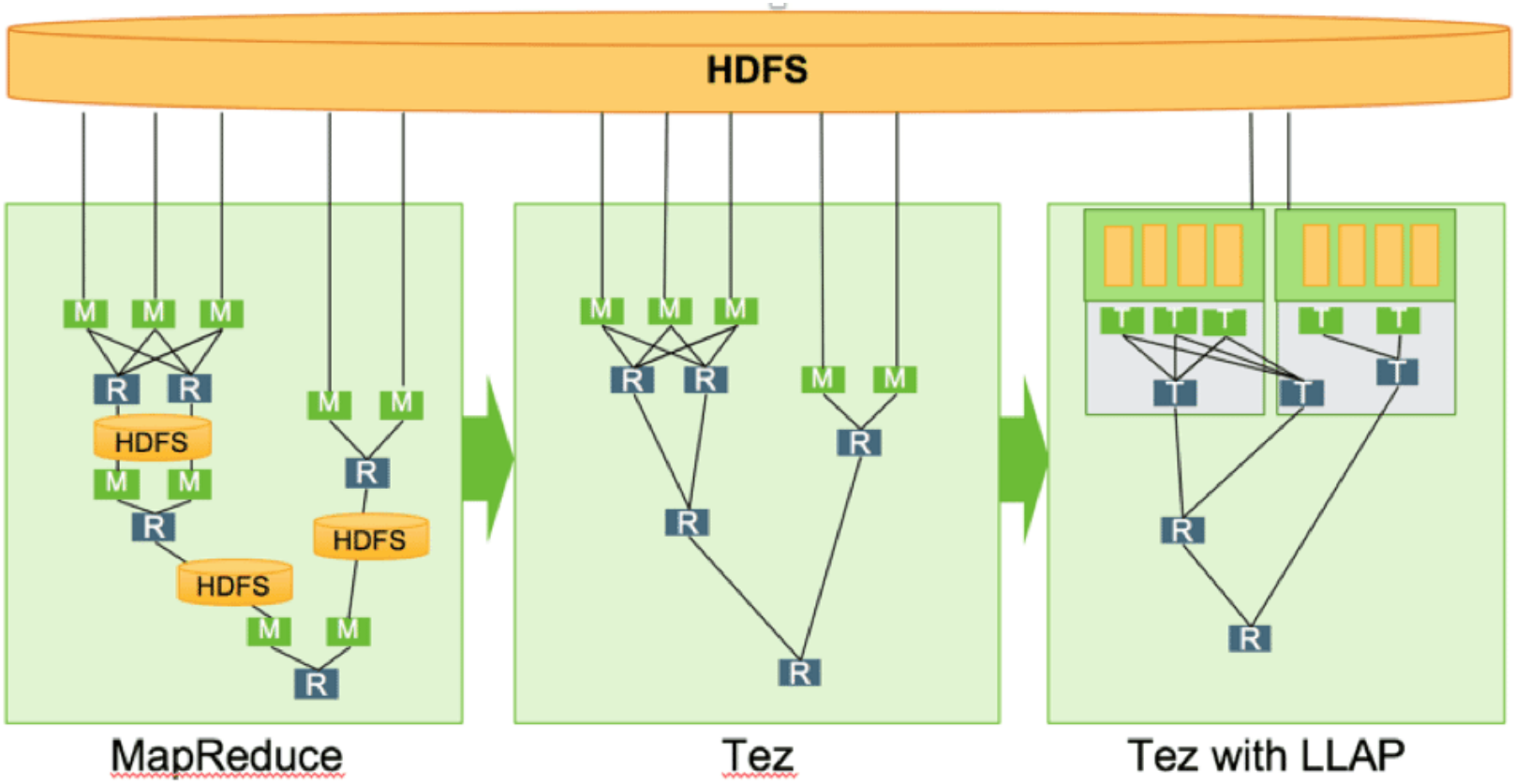
For Map machine

- Read content and prepares data from assigned portion of input data
- Feed data into Map function and saves result to local disk
- Notifies **Master** about partially completed work

For Reduce machine

- Receive notification from Master about (partially) completed Map work
- Retrieve intermediate data from Map machine via remote read
- Sorts intermediate data by key (e.g. by key word)
- Iterate over intermediate data for each unique key and sends corresponding set through Reduce function
- Add result to final output file or dataset and write it to HDFS storage.

Hadoop MapReduce improvement

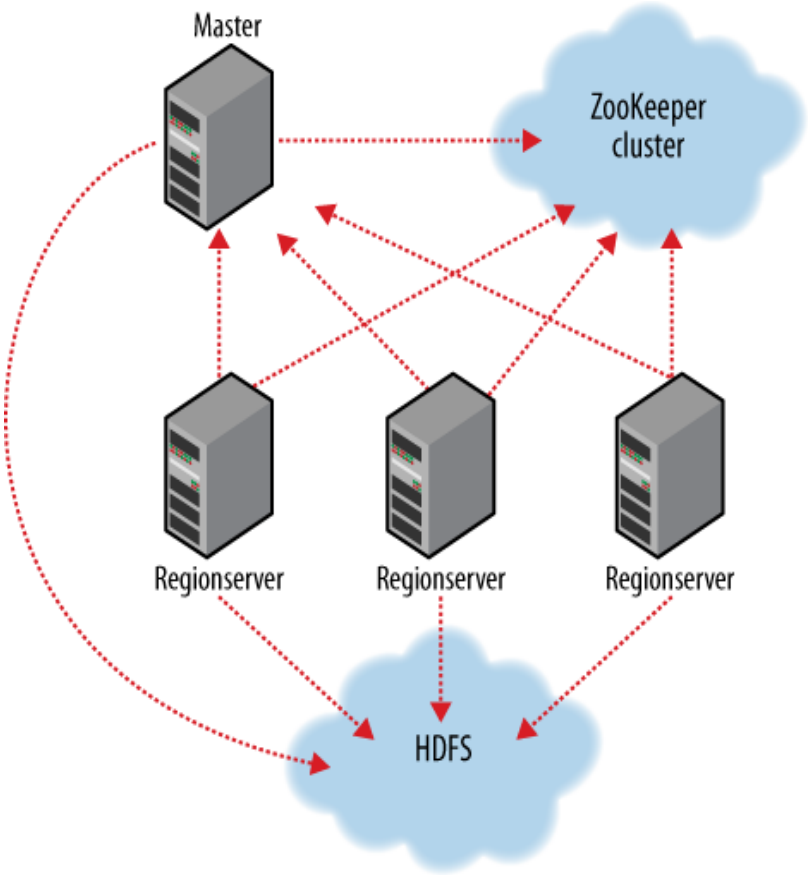


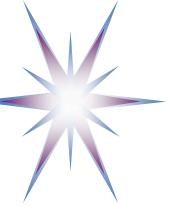
- LLAP (Live Long and Process) combines persistent query servers and optimized in-memory caching that allows Hive to launch queries instantly and avoids unnecessary disk I/O.
- Worker tasks run inside LLAP daemons, and not in containers

ZooKeeper for Managing distributed Services

Zookeeper is a highly available, scalable, distributed service for **managing consensus**, group membership, leader election, naming, configuration and coordination among distributed services and resources.

- Zookeeper is used in HBase to elect a leader and ensure that only one master is active, to manage group membership and configuration
- Ordered updates and strong persistence guarantees
  - Client will never receive old data
  - Watches for data changes
- Hierarchical namespace
  - Each znode has data and children
  - Data is read and written whole





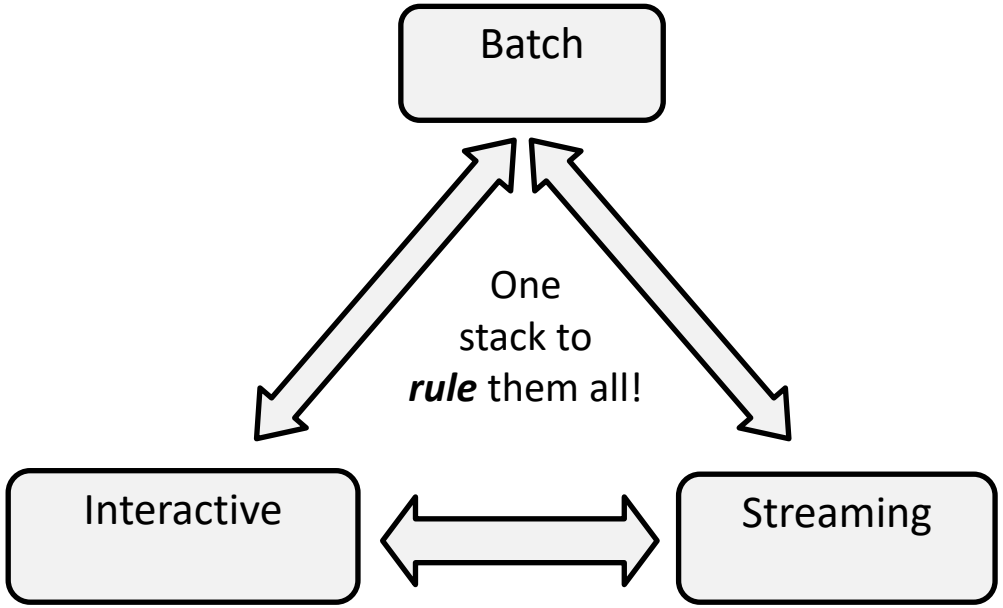
Curiously Asked Questions (CuAQ)

- Why do I need to know about Hadoop and its components?
- I head that nobody uses Hadoop, everybody are now switching to Spark
- Any other question

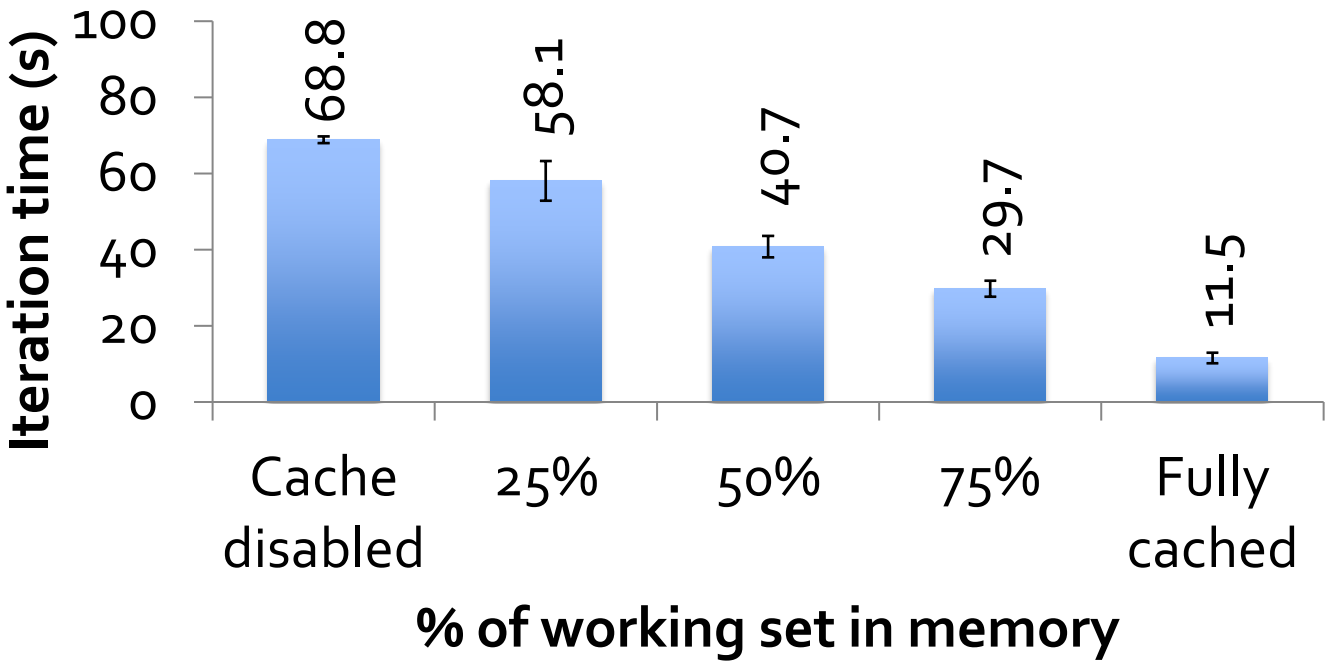
Spark and Stream Processing

- Simplicity, low latency, fault tolerance
- Expressive computing system, not limited to map-reduce model
- Uses system memory – in-memory processing
  - avoid saving intermediate results to disk
  - cache data for repetitive queries (e.g. for machine learning)
- Compatible with Hadoop

- ***Easy*** to combine ***batch***, ***streaming***, and ***interactive***e computations
- ***Easy*** to develop ***sophisticated*** algorithms
- ***Compatible*** with existing open source ecosystem (Hadoop/HDFS)



Behavior with Not Enough RAM

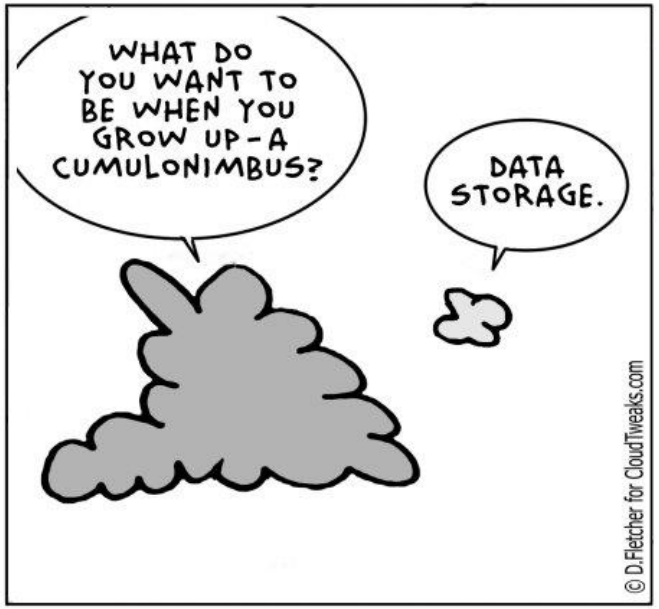


Cloud based Storage for Big Data

- Cloud storage (regular)
  - AWS Buckets and Azure Blobs - Core storage infrastructure
- Hadoop Distributed File System (HDFS)
- Data Lakes
- Large Scale Databases with controlled consistency

Cloud based Storage for Big Data

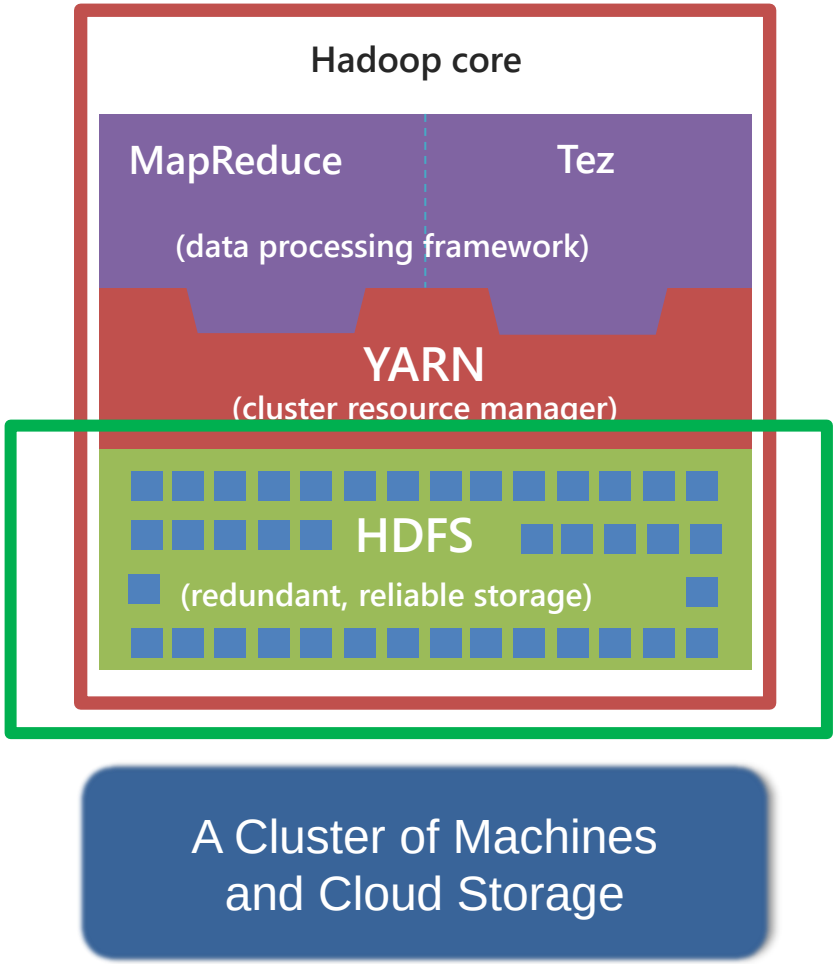
- Cloud native storage system is capable for storing large amount of data capable of existing Big Data application
- Large scale distributed file system storage
  - Can provide physical data storage for virtualized cloud storage
- Hadoop Distributed File System (HDFS)
- Data Lakes



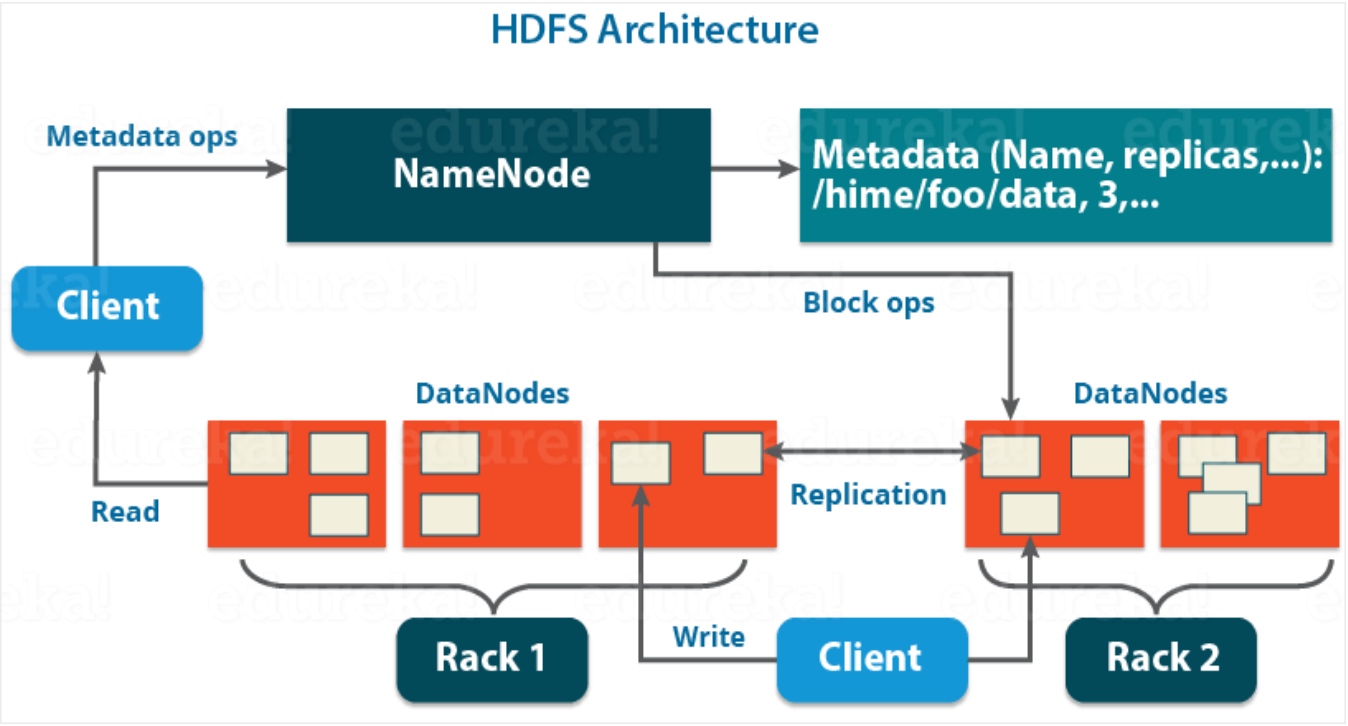
Hadoop Distributed File System (HDFS)

Hadoop is a highly reliable, distributed, and parallel programming framework for analyzing big data

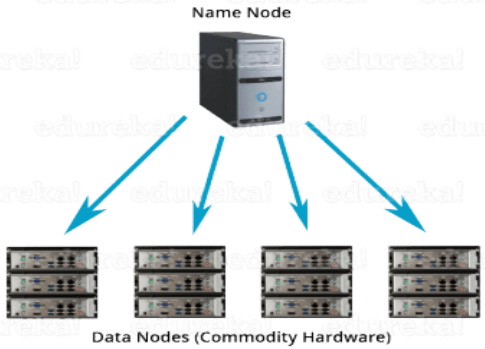
- A scalable distributed file system for large scale data analysis
- A part of the Open Source Apache Hadoop suite
  - The primary storage used by Hadoop MapReduce applications
- Can run on commodity hardware assuring high fault-tolerant
- HDFS is platform layer for Data Lakes

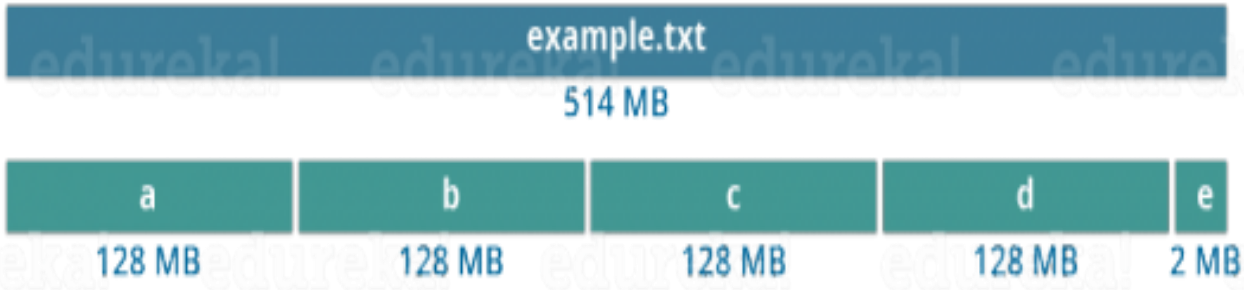


HDFS Architecture (Master – Name nodes)

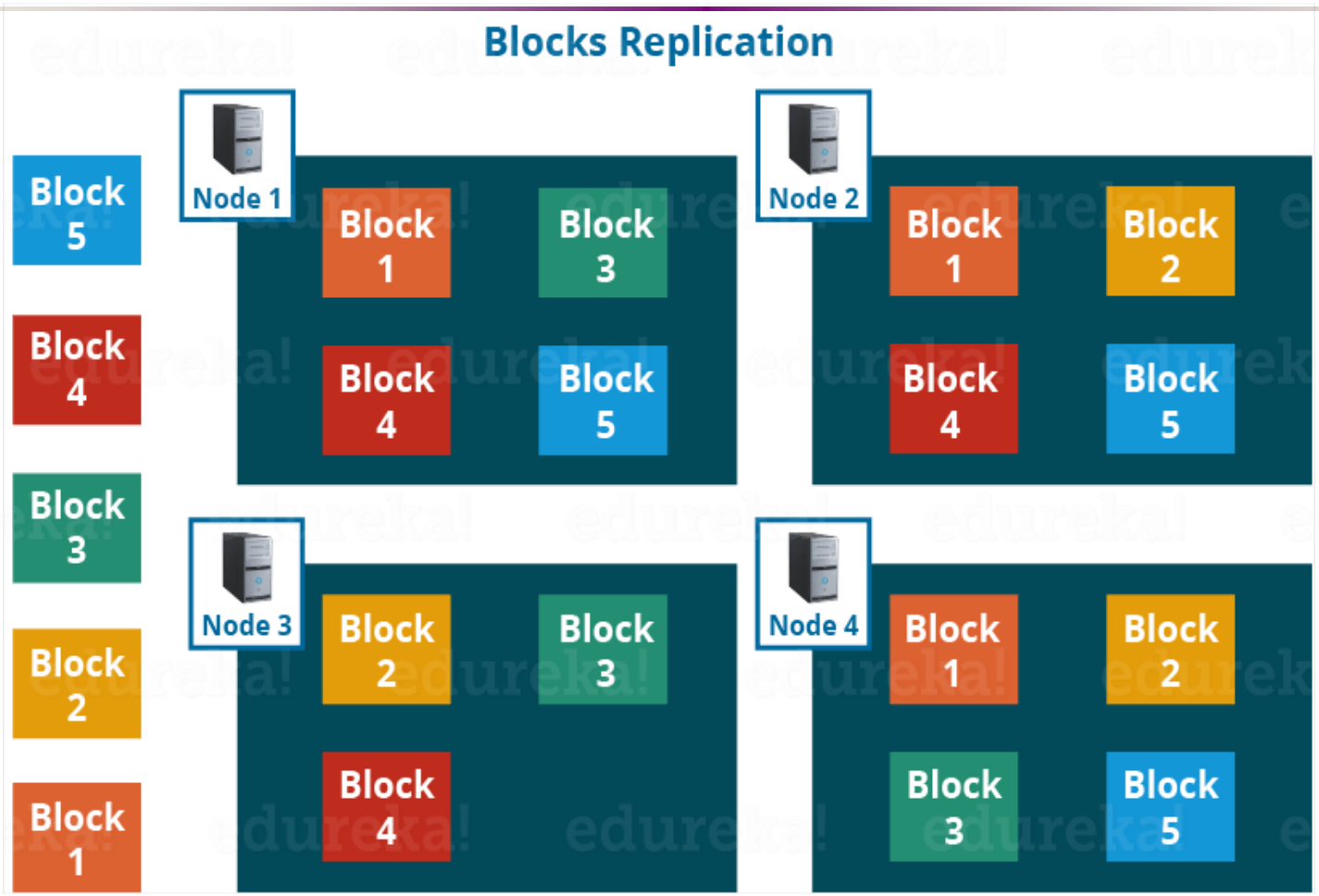


- File content is split into blocks (default 128MB, 3 replicas).
- NameNode maintains the namespace tree and the mapping of file blocks to DataNodes.
- Files and directories are represented on the NameNode by *inodes* (permissions, modification and access times, namespace and disk space quotas).
- Namespace is a hierarchy of files and directories.

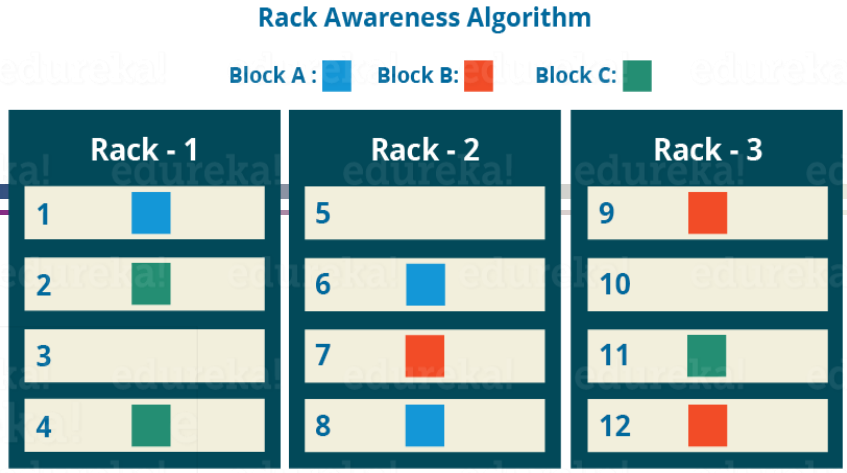


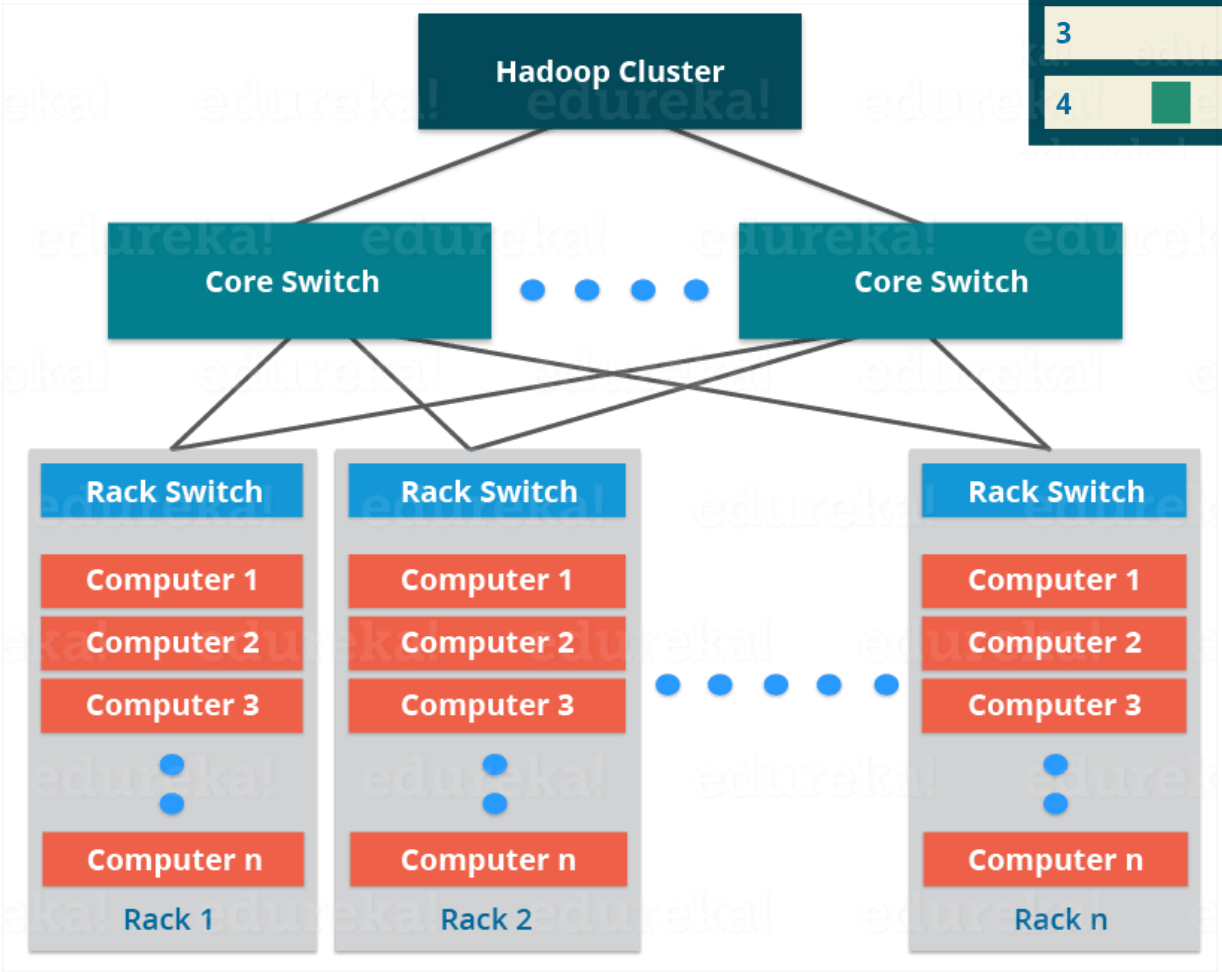


HDFS Replication Management



Rack Awareness





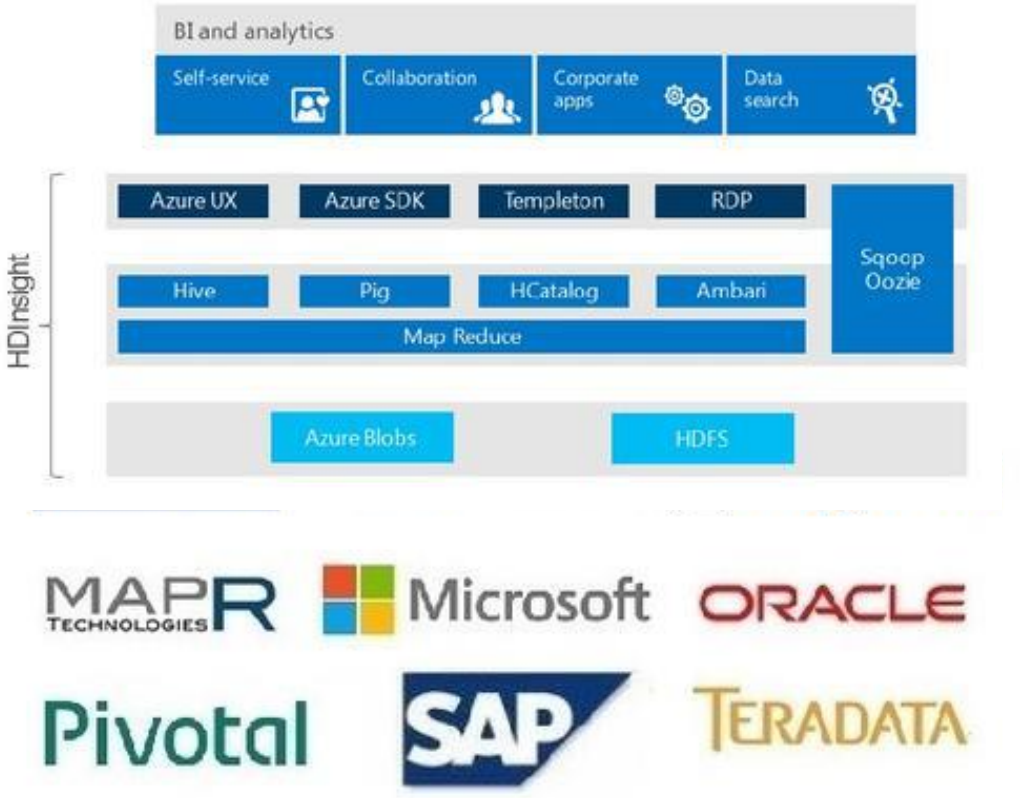
- HDFS racks are designed with advanced network performance and design/wiring

Platforms for Big Data and Data Analytics

- Hadoop and Spark based
- Traditional database and warehouse
- **AWS Example**
  - **Analytical DBMS:** Amazon Redshift service (based on ParAccel engine); Amazon Relational Database Service.
  - **In-memory DBMS:** None. Third-party options on AWS include Altibase, SAP Hana, and ScaleOut.
  - **Hadoop distributions:** Amazon Elastic MapReduce. Third-party options include Cloudera and MapR.
  - **Stream-processing technology:** Amazon Kinesis.

[ref] Top 16 Big Data Platforms, 2016 https://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609

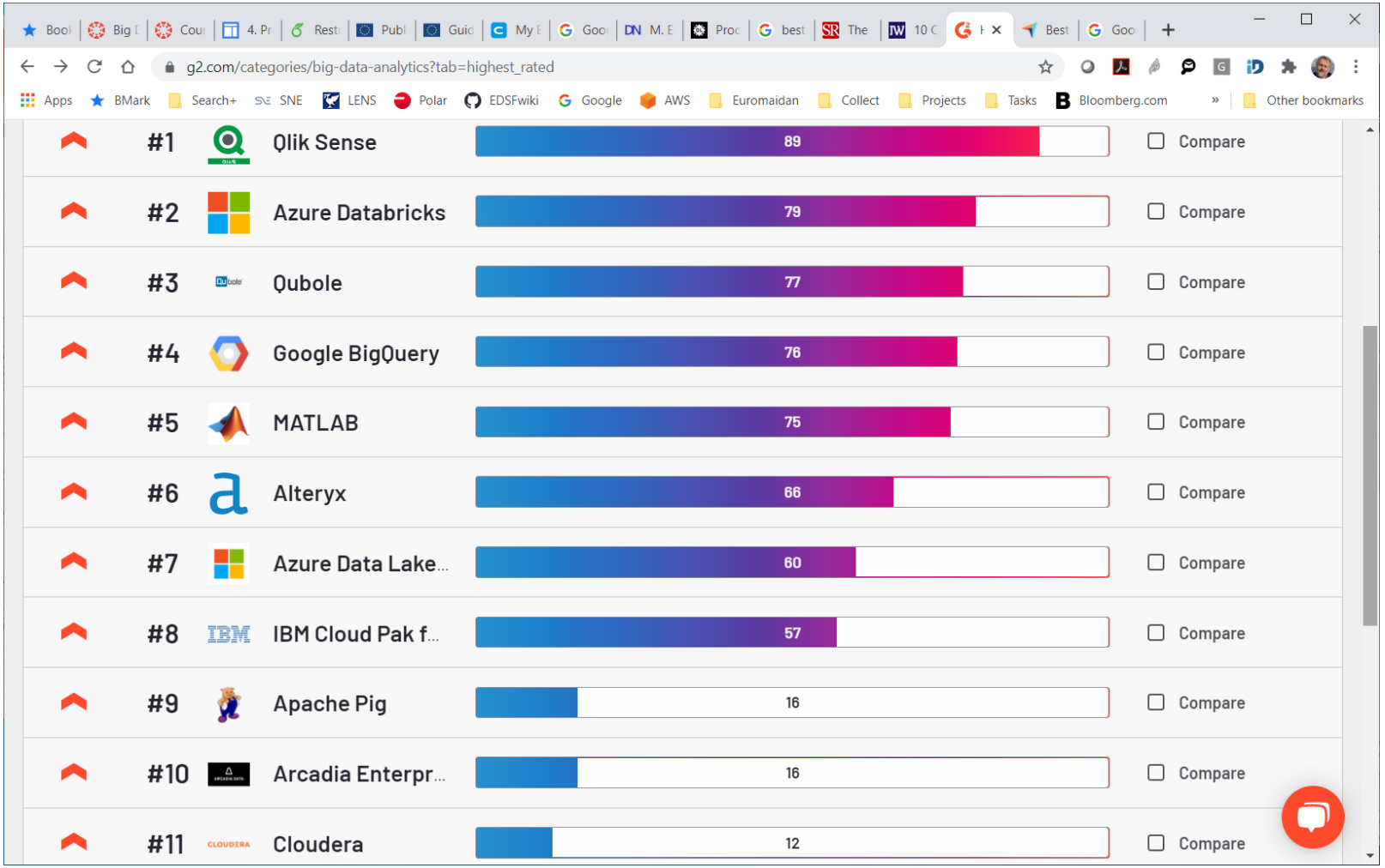
Platforms for Big Data and Data Analytics



The Top 11 **Big Data Analytics Software**

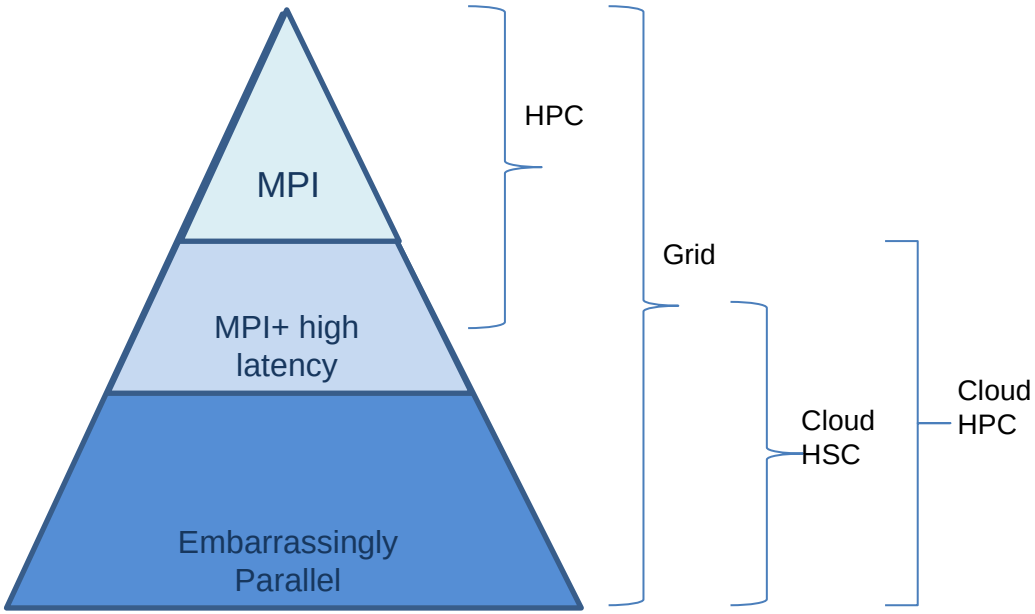
- Hadoop and Spark based
- Traditional database and warehouse
- **Microsoft Example**
  - **Analytical DBMS:** SQL Server 2012 Parallel Data Warehouse (PDW).
  - **In-memory DBMS:** SQL Server 2014 In-Memory OLTP (option available with SQL Server 2014, set for release by second quarter of 2014).
  - **Stream-processing technology:** Microsoft StreamInsight.
  - **Hadoop distribution:** HDInsight/Windows Azure HDInsight Service (based on Hortonworks Data Platform).
  - **Hardware/software systems:** Dell Parallel Data Warehouse Appliance, HP Enterprise Parallel Data Warehouse Appliance.

The Top 11 Big Data Analytics Software 2020 [ref]



[ref] https://www.g2.com/categories/big-data-analytics?tab=highest_rated

HPC and Cloud



Relations between HPC, Cloud HSC, Cloud HPC, and Grid computing models

- Message Passing Interface (MPI) and clustering requiring low latency network and high performance I/O processes
- MPI + high latency adopt using MPI with the distributed computing resources
- Embarrassingly Parallel Problem (EPP) computing benefits from using distributed computing resources and possibility to scale dynamically

Data Lakes - Definition

Data Lake allows an organization to store all of their data, structured and unstructured, in one, centralized repository.

- Since data can be stored as-is, there is no need to convert it to a predefined schema and you no longer need to know what questions you want to ask of your data beforehand.

A Data Lake should support the following capabilities:

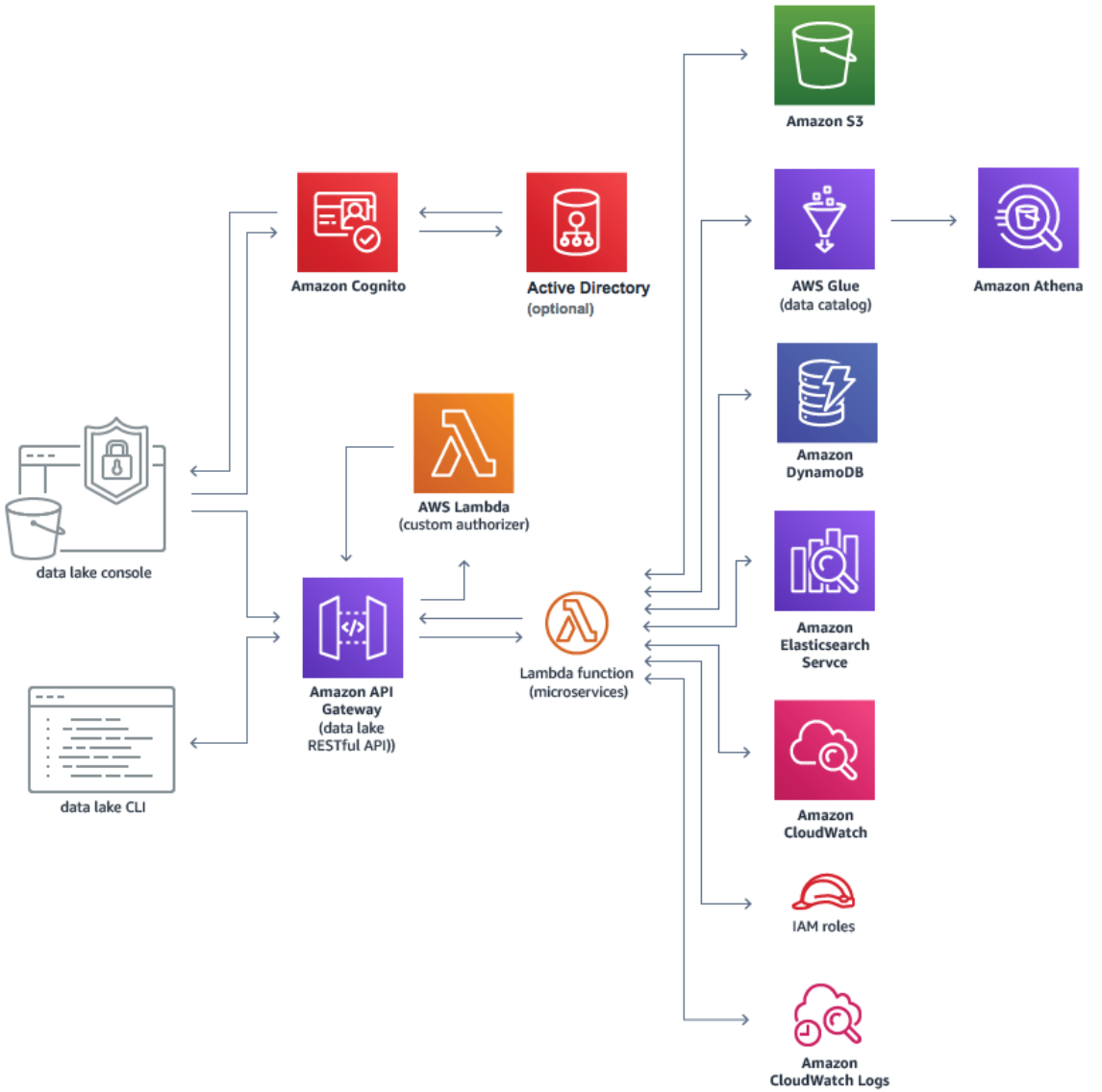
- Collecting and storing any type of data, at any scale and at low costs
- Securing and protecting all of data stored in the central repository
- Searching and finding the relevant data in the central repository
- Quickly and easily performing new types of data analysis on datasets
- Querying the data by defining the data's structure at the time of use (schema on read)

Data Lake Layers

- **Raw data layer** – Raw events are stored for historical reference. Also called staging layer or landing area
- **Cleansed data layer** – Raw events are transformed (cleaned and mastered) into directly consumable data sets. Aim is to uniform the way files are stored in terms of encoding, format, data types and content (i.e. strings). Also called conformed layer
- **Application data layer** – Business logic is applied to the cleansed data to produce data ready to be consumed by applications (i.e. DW application, advanced analysis process, etc). Also called workspace layer or trusted layer

Still need data governance so your data lake does not turn into a data swamp!

AWS Data Lake Architecture



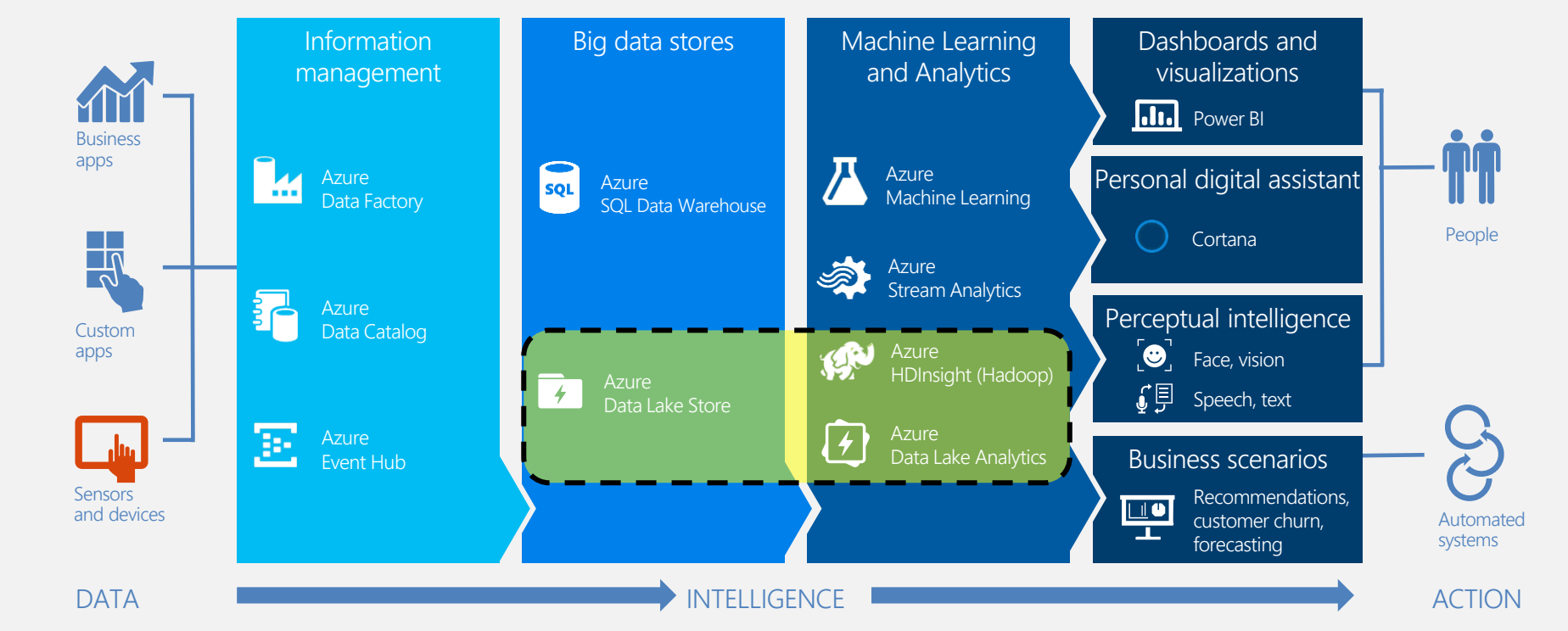
Components

- Amazon S3
- AWS Glue and Amazon Athena
- Amazon DynamoDB
- Amazon Elasticsearch Service (Amazon ES)
- Amazon CloudWatch
- Amazon API Gateway
- Amazon Cognito and Active Directory
- Amazon Lambda

- Can be implemented using accompanying CloudFormation Template

https://aws.amazon.com/solutions/implementations/data-lake-solution/

Azure Data Lake

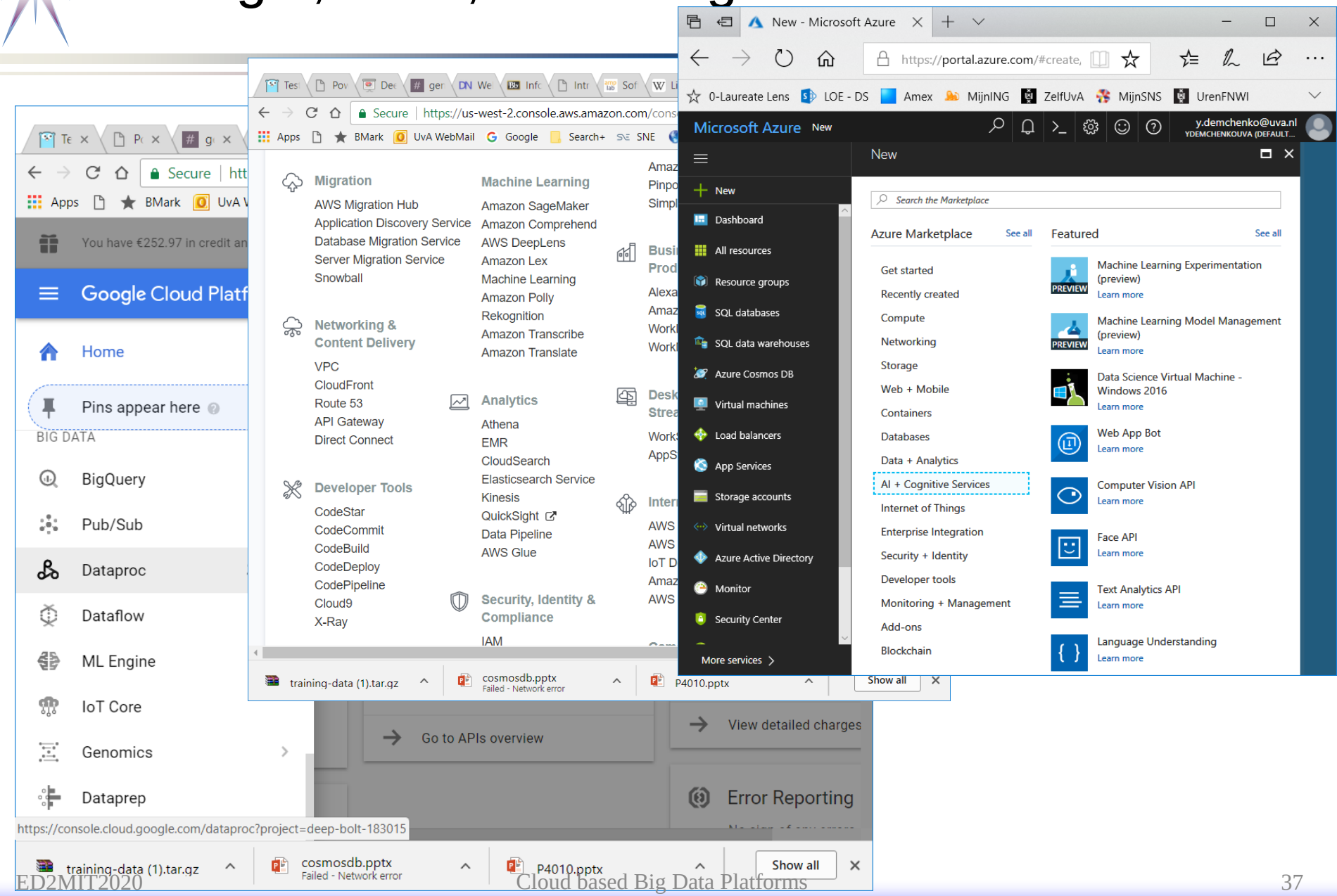


- Azure Data Lake Analytics is a part of Cortana Analytics Suite

Cloud based Big Data Platforms and Solutions

- Amazon Web Services (AWS)
- Google Cloud Platform (GCP)
- Microsoft Azure

Google, AWS, Azure Big Data Stacks



AWS Cloud Big Data Services

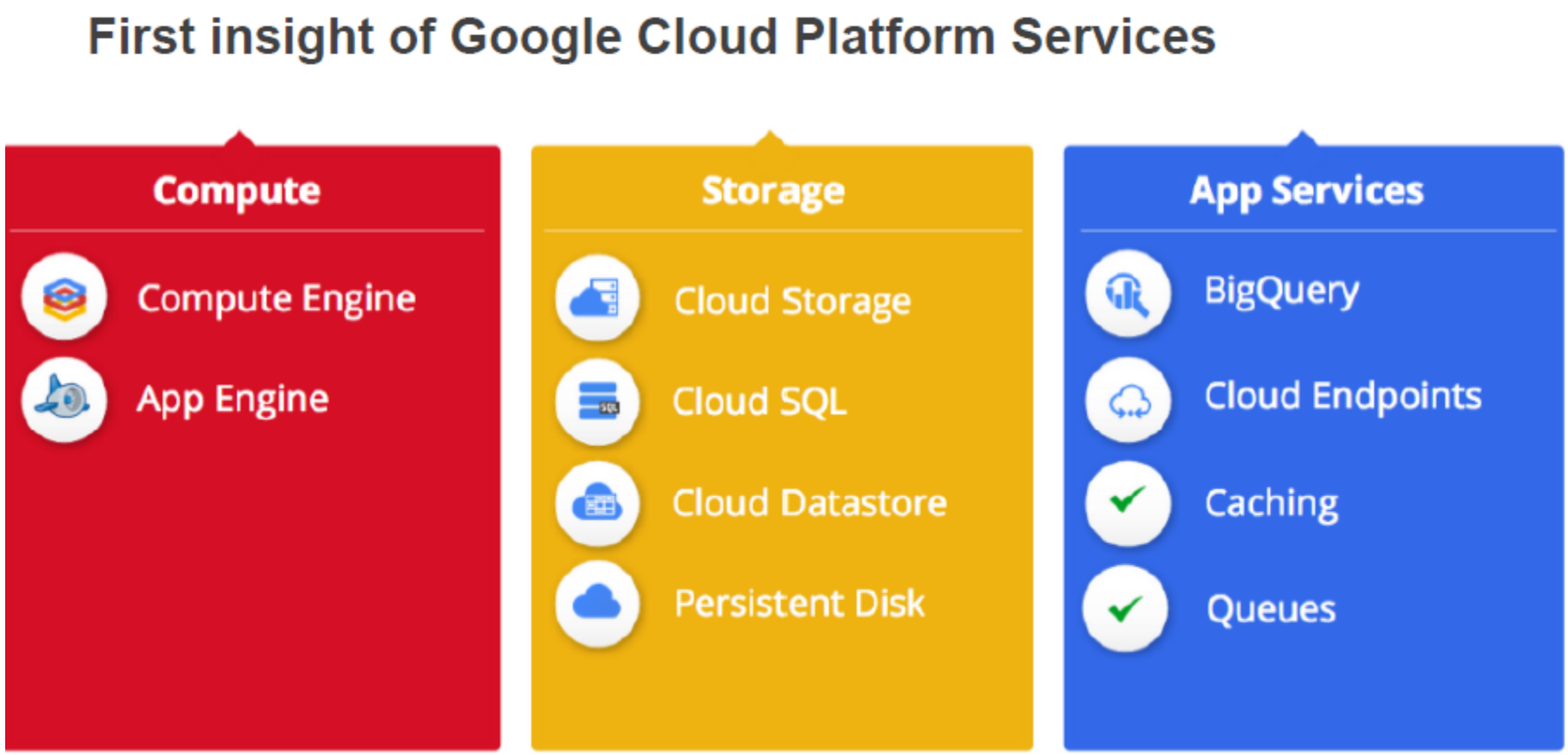
AWS Cloud offers the following services and resources for Big Data processing:

- EC2 Virtual Machine (VM) instances for HPC optimized for computing (with multiple cores) and with extended storage for large data processing.
- **Amazon Elastic MapReduce (EMR)** provides the Hadoop framework on Amazon EC2 and offers a wide range of Hadoop related tools.
- **Amazon Kinesis** is a managed service for real-time processing of streaming big data (throughput scaling from megabytes to gigabytes of data per second and from hundreds of thousands different sources).
- **Amazon DynamoDB** highly scalable NoSQL data stores with sub-millisecond response latency.
- Amazon Aurora scalable relational database.
- Amazon Redshift fully-managed petabyte-scale data warehouse in cloud at cost less than $1000 per terabyte per year.
- Amazon Machine Learning
- Machine Learning (Artificial Intelligence) based services (Lex, Translate, Recognition, etc.)

Google Cloud Platform (GCP)

- Compute
  - AppEngine
  - Google Functions (serverless with Node.js)
- Storage: Static, sharing, backup, for applications and computation
  - Cloud Spanner SQL database
- Big Data
  - BigQuery – Hadoop Data Warehouse
- Machine Learning services
  - Translate
  - Prediction
- Cloud endpoints

Google Cloud Platform (GCP) structure



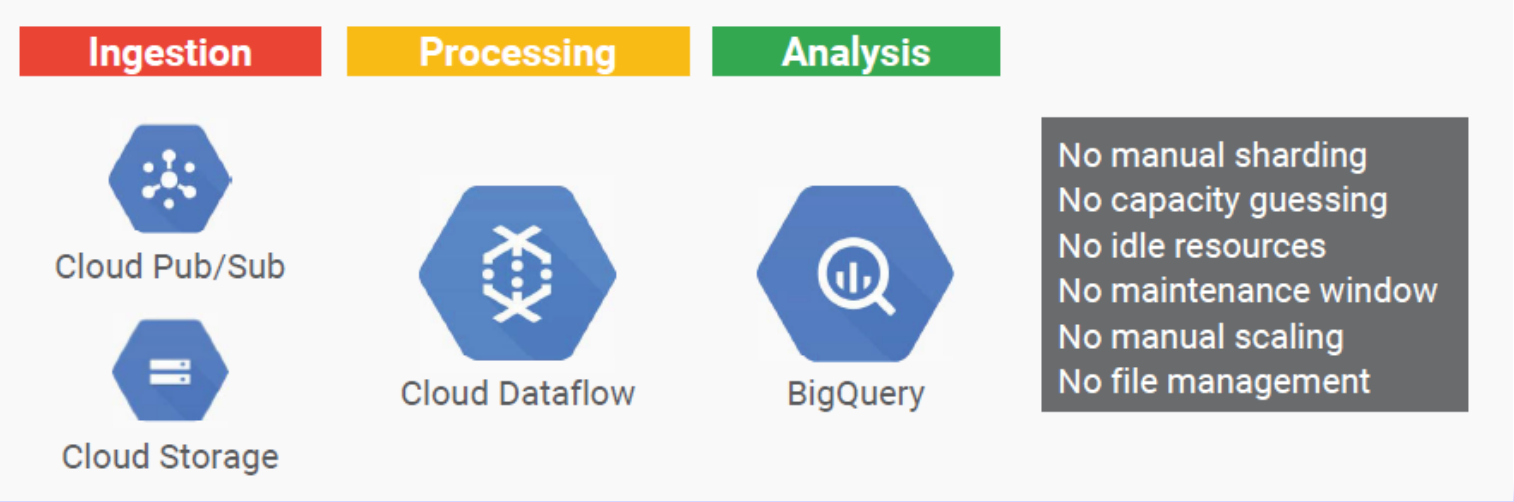
Machine Learning Focus

- Machine Learning embedded across most products
- Multiple Tensorflow ML models in use
  - Portable TensorFlow models
- Key models exposed via APIs (Democratizing Machine Learning)
  - Cloud Video Intelligence API
  - Cloud Vision API
  - Cloud Natural Language API
  - Cloud Translation API
  - Cloud Speech API

- Acquired Kaggle in 2017 - Data Science Enthusiasts

Google Machine Learning

- Support all stages of ML workflow
- Dataprep: Serverless platform for all stages of the analytics data lifecycle

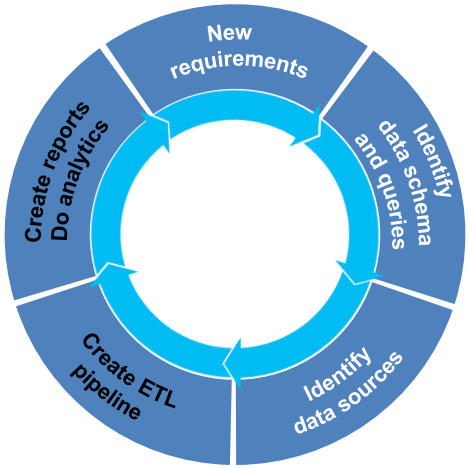


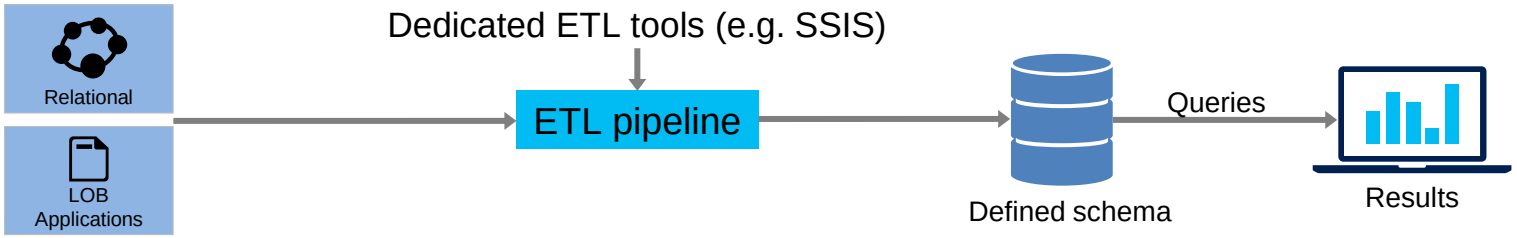
Microsoft Azure Big Data Services

- New Big Data data-centric thinking
- Azure Data Lakes
- HDInsight
- MLOps

Traditional business analytics process

1. Start with end-user requirements to identify desired reports and analysis
2. Define corresponding database schema and queries
3. Identify the required data sources
4. Create a Extract-Transform-Load (ETL) pipeline to extract required data (curation) and transform it to target schema (‘*schema-on-write*’)
5. Create reports, analyze data

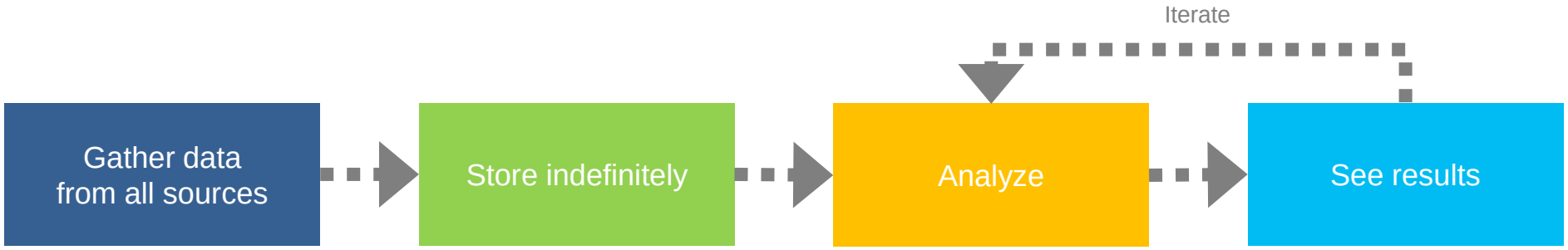




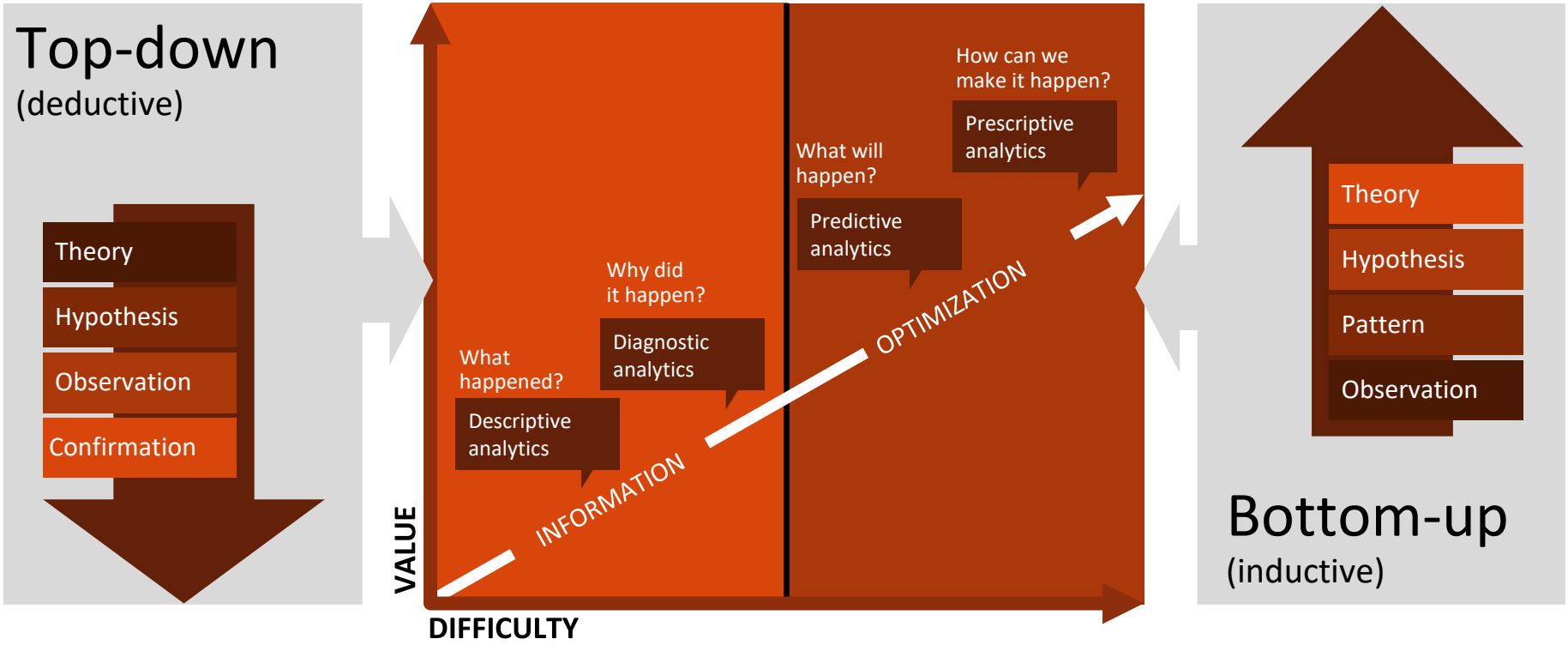
All data not immediately required is discarded or archived

New Big Data thinking: All data has value

- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time *(schema-on-read).*
- Apps and users interpret the data as they see fit

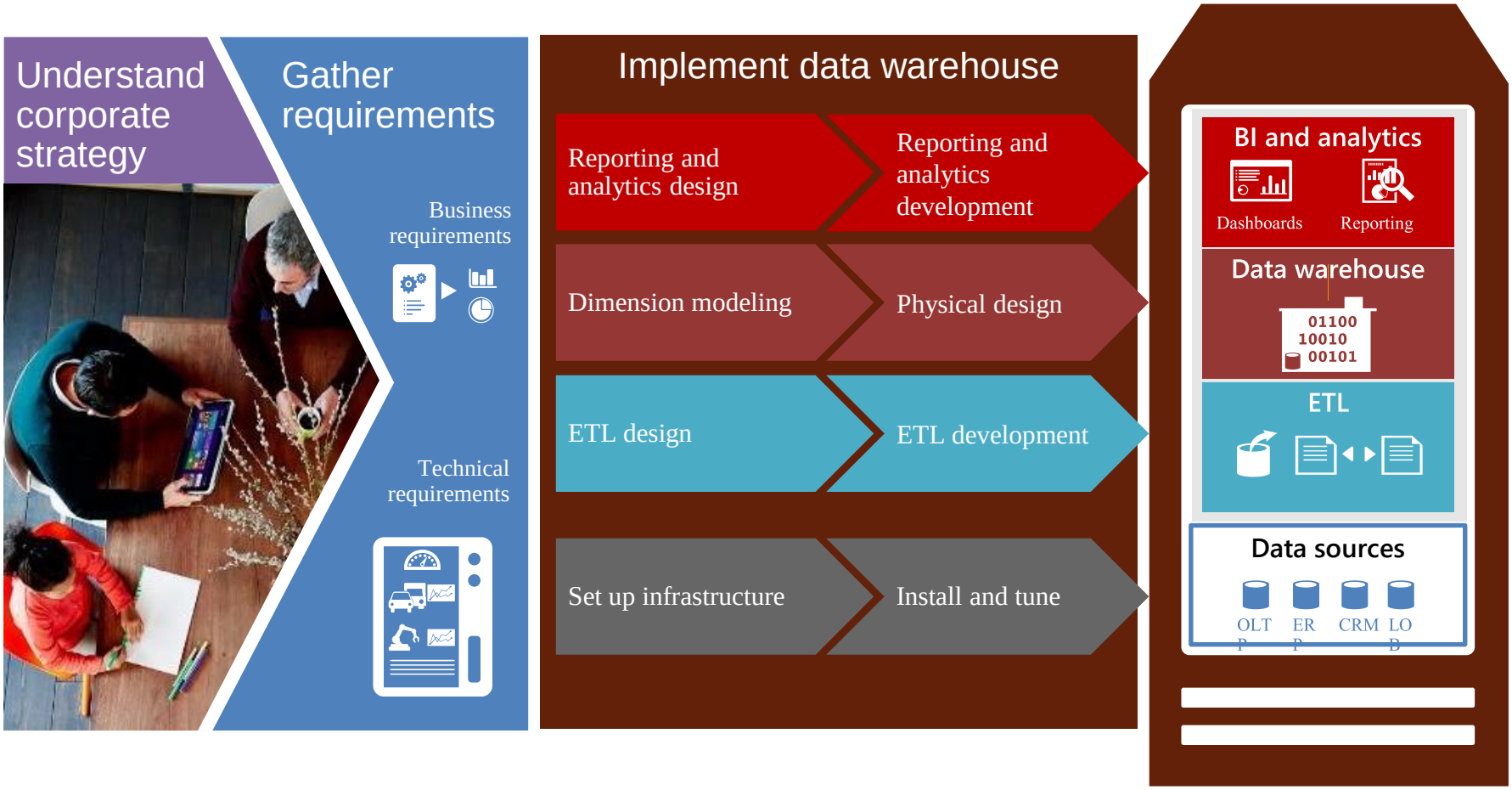


Two approaches to information management for analytics: Top-down and bottom-up



- Top-down: Start with the data model creation
- Bottom-up: Collect data, discover patterns and extract value

Data Warehousing uses a top-down approach



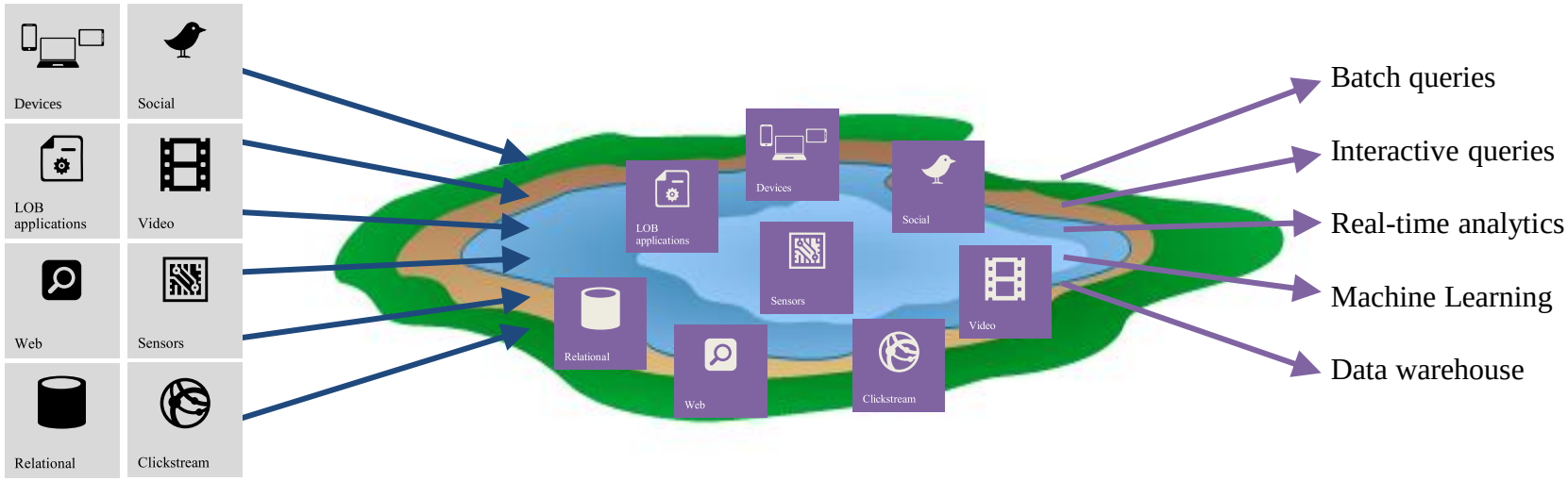
- ETL – Extract – Transfer – Load
- ELT – Extract – Load - Transfer

The Data Lake uses a bottom-up approach

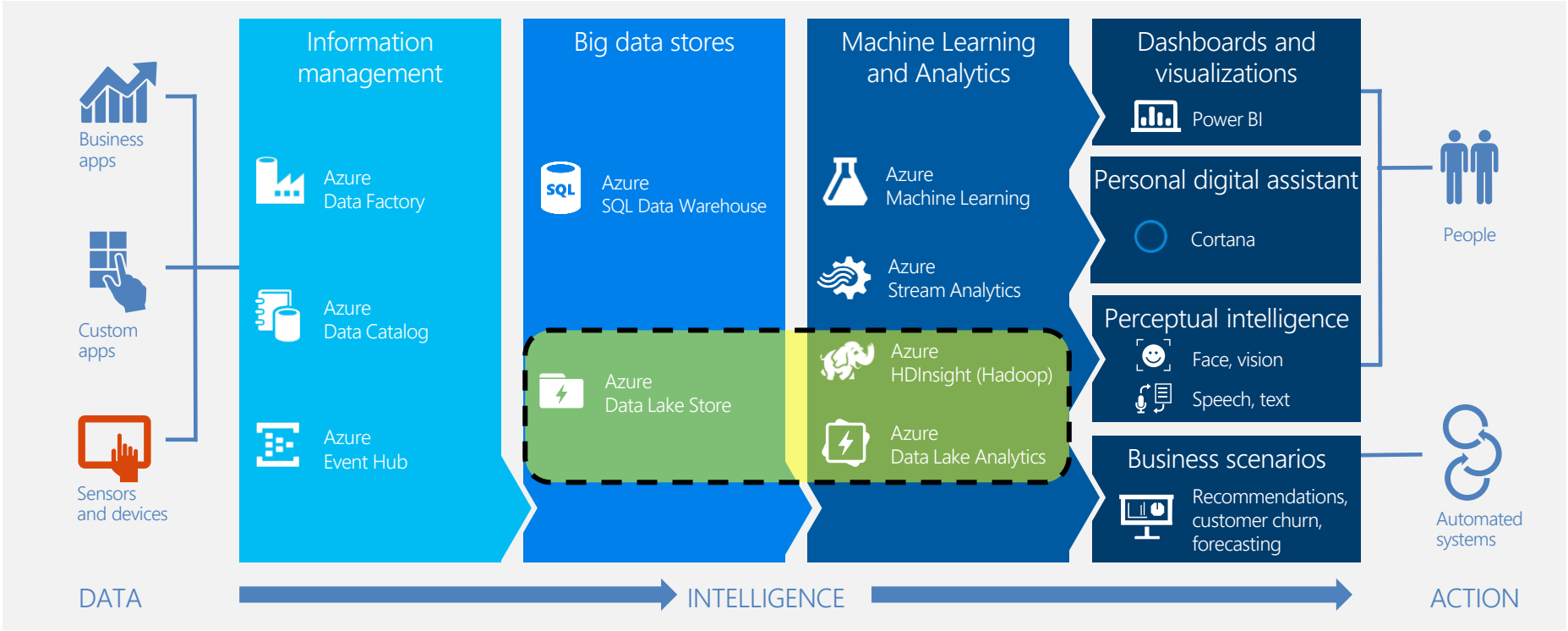
Ingest all data
regardless of requirements

Store all data
in native format without schema definition

Do analysis
using analytic engines like Hadoop



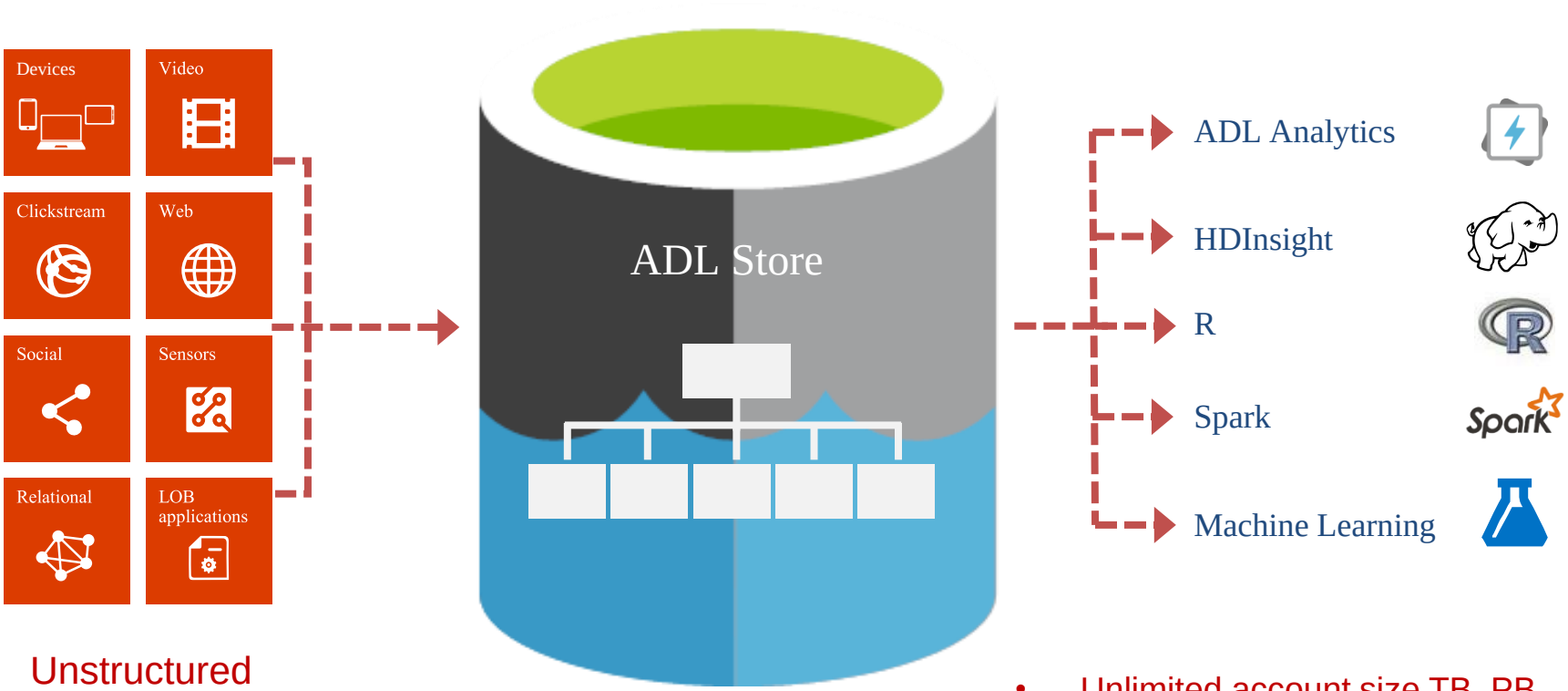
Azure Data Lake



- Part of Cortana Analytics Suite

What is Azure Data Lake (ADL) Store?

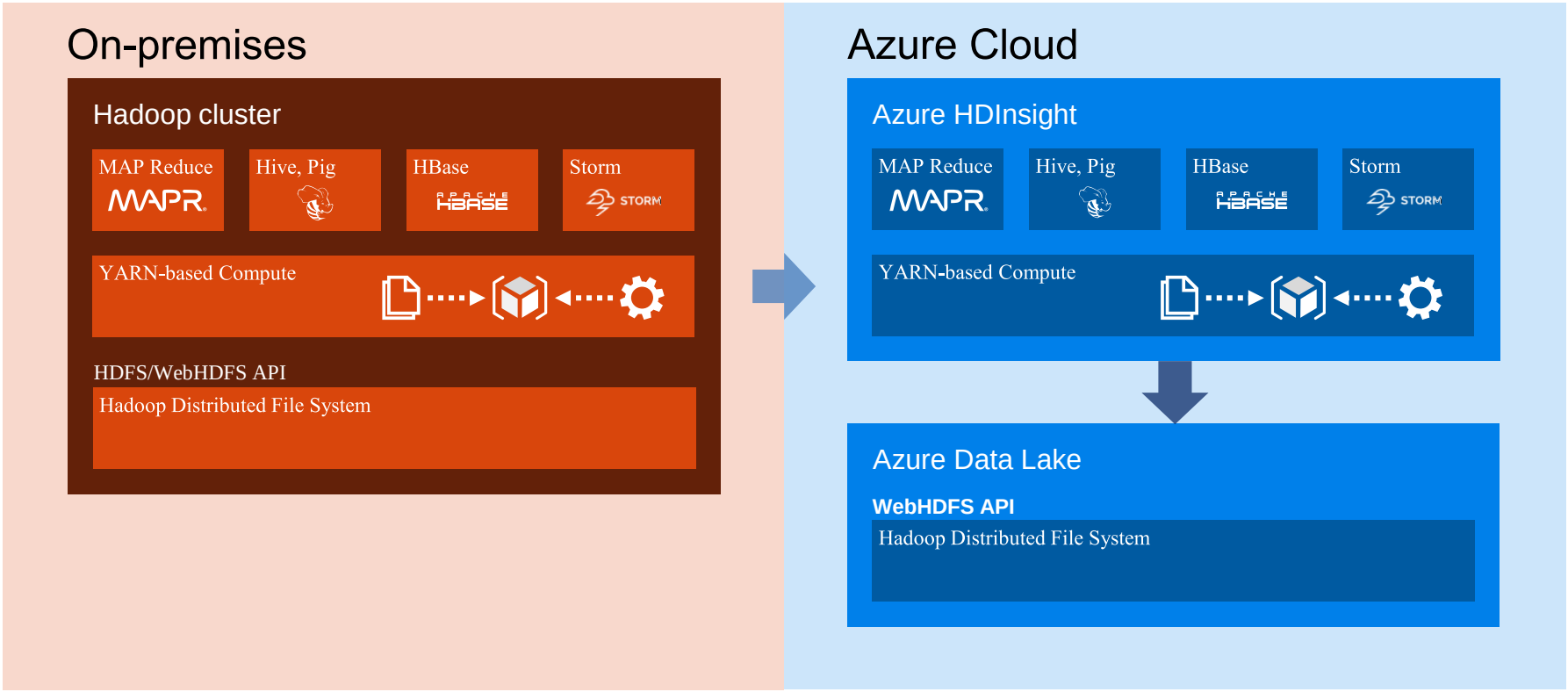
- A highly scalable, distributed, parallel file system in the cloud specifically designed to work with multiple analytic frameworks



- Unstructured
- Semi-structured
- Structured

- Unlimited account size TB, PB
- Individual files size from gigabytes to petabytes
- No limits to scale

Hybrid Data Lake Model in Azure



- Azure Stack – fully functional cloud software, is installed on premises
- It communicate with Azure cloud for data storage, processing and data analytics

Azure HDInsight – What is it?

A standard Apache Hadoop distribution offered as a managed service on Microsoft Azure

- Based on Hortonworks Data Platform (HDP)
- Provisioned as clusters on Azure that can run on Windows or Linux servers
- Offers capacity-on-demand, pay-as-you-go pricing model
- Integrates with:
  - Azure Blob Storage and Azure Data Lake Store for Hadoop File System (HDFS)
  - Azure Portal for management and administration
  - Visual Studio for application development tooling

In addition to the core, HDInsight supports the Hadoop ecosystem

Other Big Data Platforms

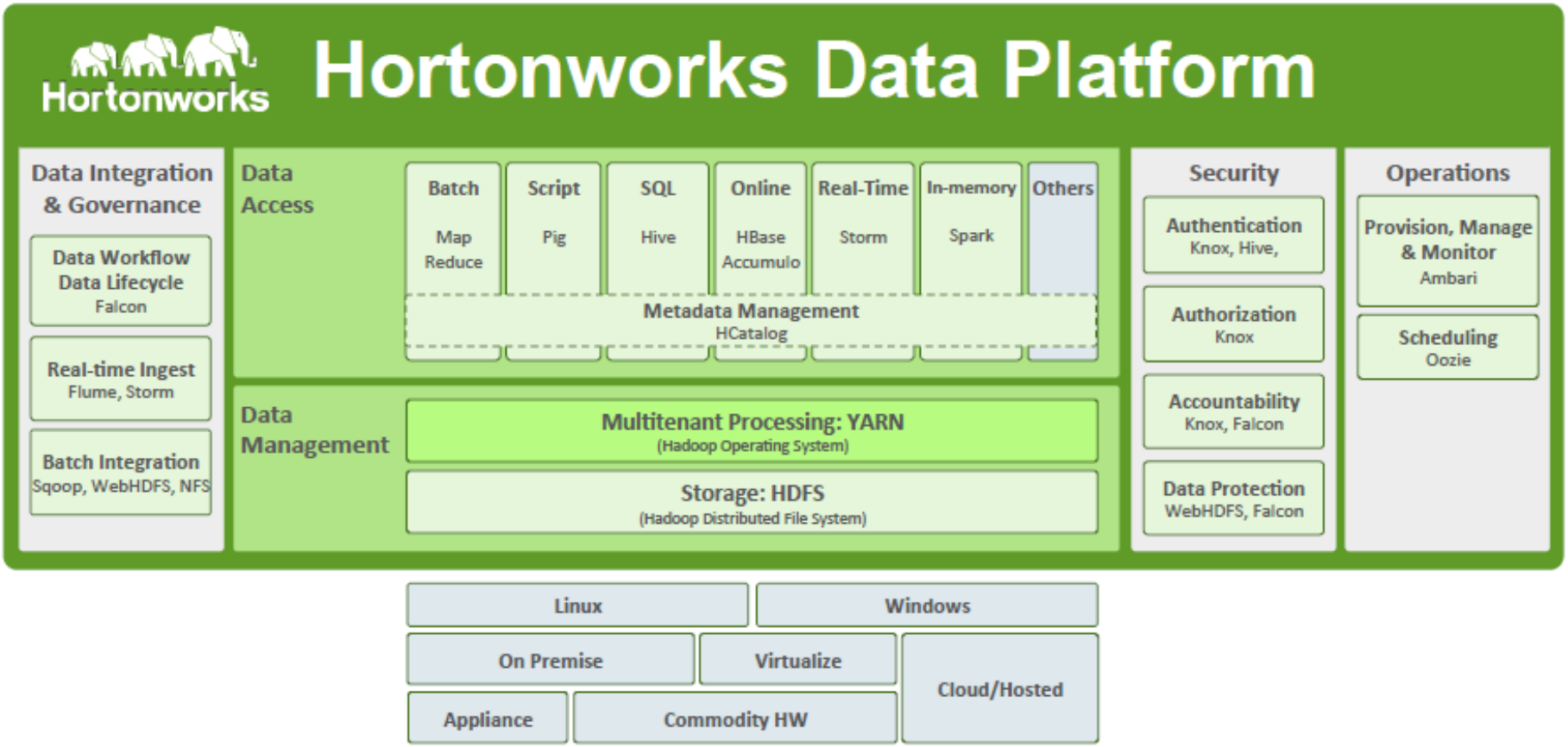
- Hortonworks – Recently merged with/into Cloudera
- Cloudera
- Oracle

- All three platforms has versions for laptop/destop installation

Hortonworks Data Platform (HDP)
http://hortonworks.com/

- HDP delivers a single integrated Hadoop platform for enterprises
  - Provides a data platform for multi-workload data processing across an array of processing methods including batch and interactive to real-time
  - Supports key capabilities of an enterprise data platform: Governance, Security and Operations
- YARN and Hadoop Distributed Filesystem (HDFS) are the core components of HDP
  - Allows creating multi-tenant data analytics applications
- HDP runs natively on Linux and Windows OS
  - HDP provides the basis for Azure HDInsight Service meaning complete portability of data is retained on-premise and in the cloud
  - Available in integrated hardware from Teradata
- Hortonworks provides a simple starters solution Hadoop Sandbox
  - Hortonworks Sandbox is a single-node implementation of Hadoop based on the Hortonworks Data Platform that includes all the typical components found in a Hadoop deployment

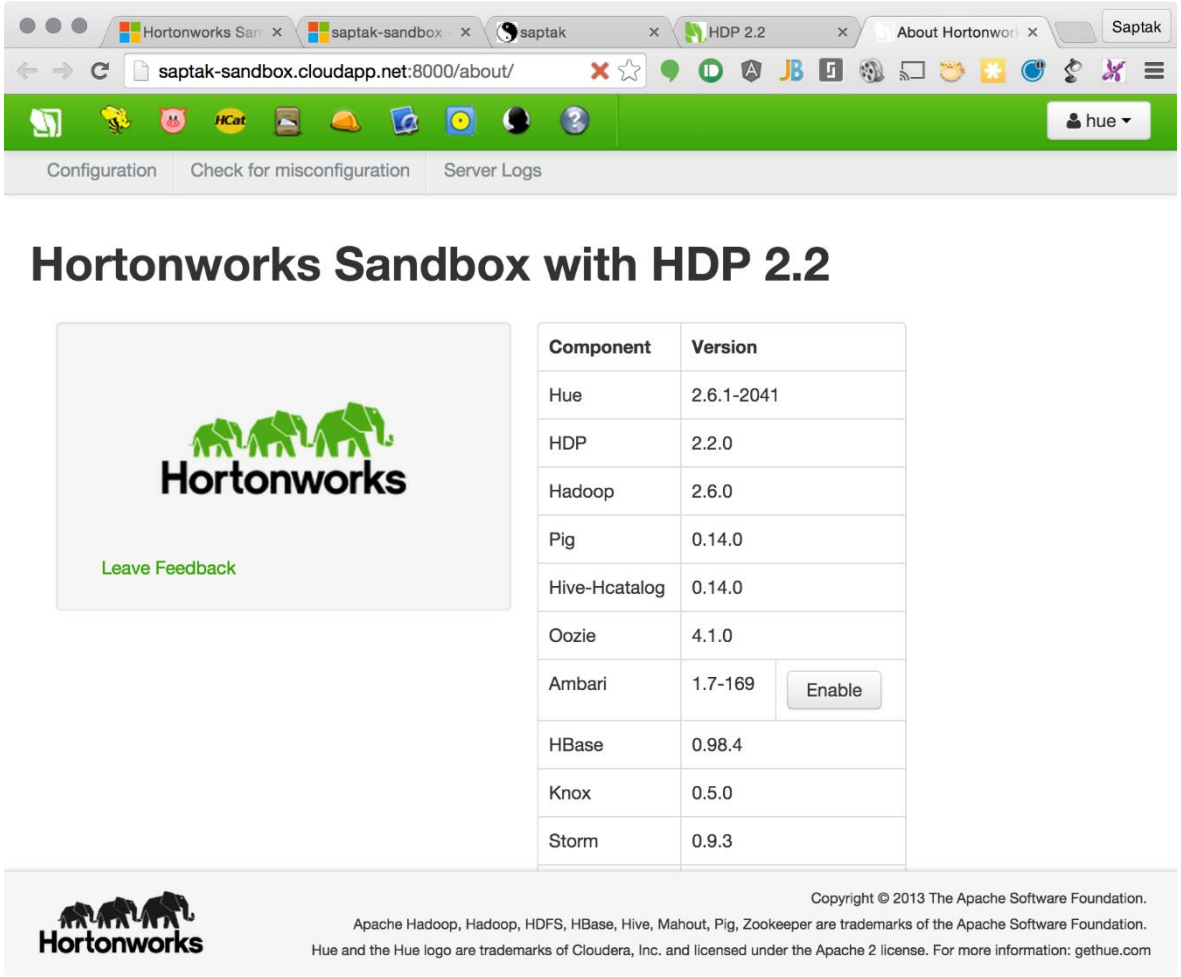
Hortonworks Data Platform Architecture [ref]



- HDP includes the most recent developments of the Open Source Hadoop suite
- Can run on Linux and on Windows OS
- Can be deployed on premises on dedicated cluster and on cloud as a hosted application

[ref] http://hortonworks.com/hdp /

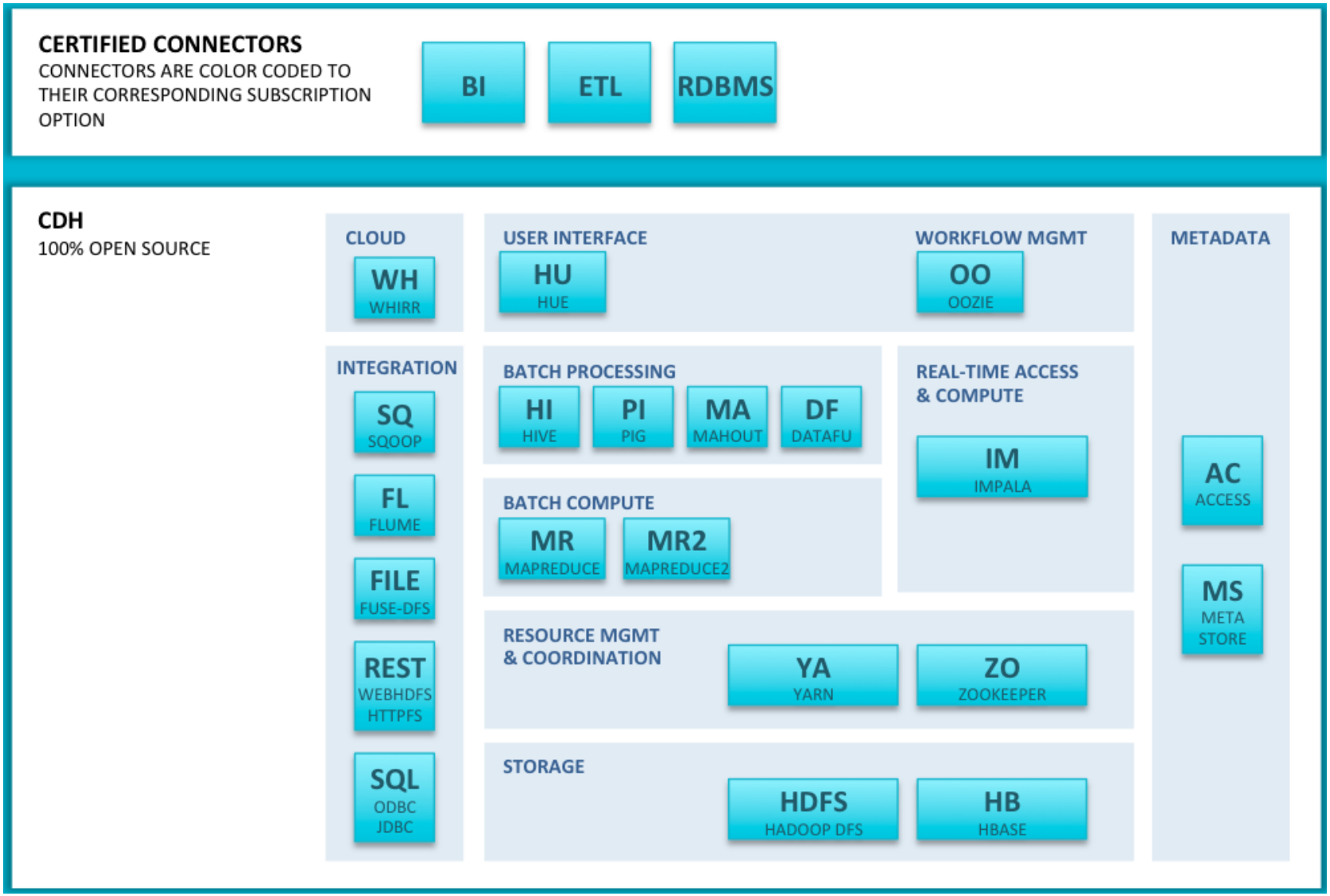
Hortonworks Sandbox VM – Standalone



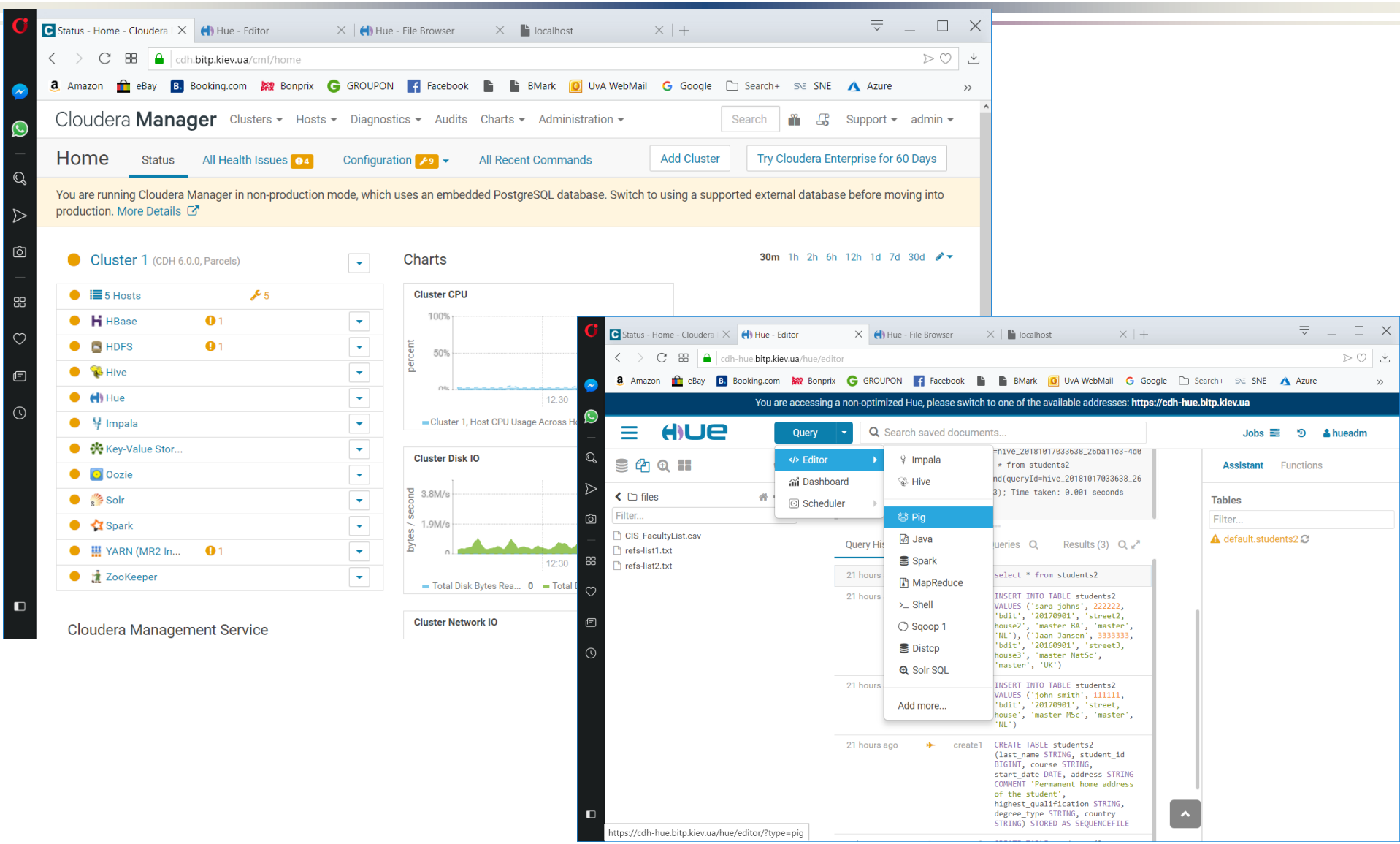
Simple starter solution Hadoop Sandbox

- Hortonworks Sandbox is a single-node implementation of Hadoop based on the Hortonworks Data Platform
- Includes all the typical components found in a Hadoop deployment

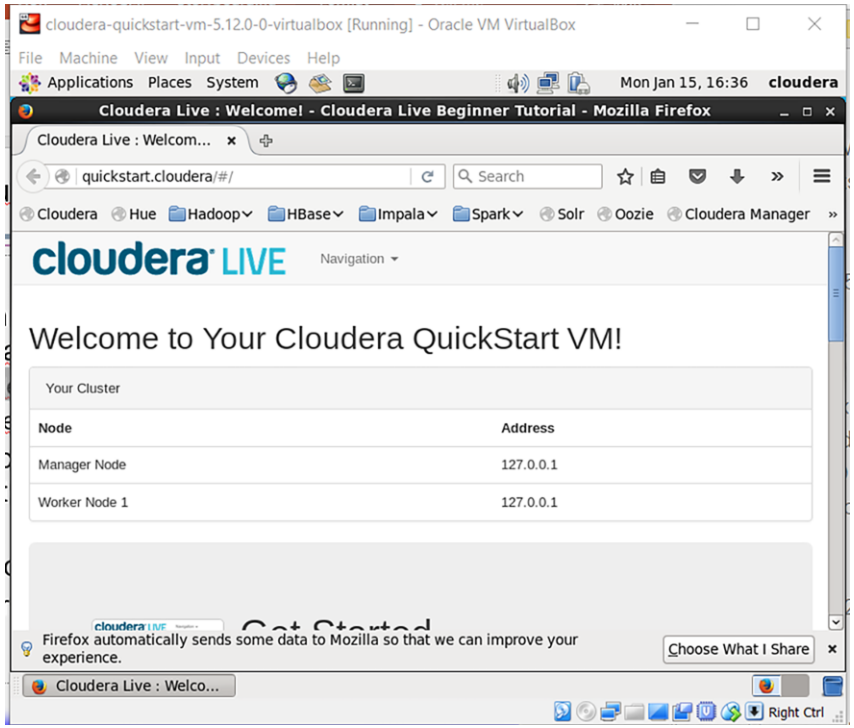
Cloudera Hadoop Cluster Architecture



Cloudera Hadoop cluster on cloud



Cloudera Quickstart VM for VirtualBox



Accounts

- Once you launch the VM, you are automatically logged in as the cloudera user. The account details are:
  - username: cloudera
  - password: cloudera
- The cloudera account has sudo privileges in the VM. The root account password is cloudera.
- The root MySQL password (and the password for other MySQL user accounts) is also cloudera.
- Hue and Cloudera Manager use the same credentials.

Mastering Big Data Tools and Services

- Educational resources
- Free/trial accounts

Environment and tools

- AWS educational class
  - All individual accounts with 50$ credit
  - Limits and quotas for AWS resources applied
  - Consider applying for individual AWS educational account
- Note: Be cost aware when using cloud resource
  - Check price of resources: instances, services, transactions, memory use
  - Try to pool together for the group activity
    - Balance individual practice work and group project work
- Your personal development environment
  - Cloudera Starter VirtualBox, Eclipse, Visual Studio Code, Python
  - Github for joint project development

Educational materials

- Available on Canvas
  - Recommended literature and links to DevOps related resources and tools
  - Some DevOps related papers and studies on Canvas
- Search web, use Wikipedia and other *pedia as a first step, always check sources
- Online courses and tutorials
  - AWS learning resources
    https://aws.amazon.com/training/self-paced-labs/
    https://aws.amazon.com/getting-started/
  - Microsoft Learn https://docs.microsoft.com/en-us/learn/
    - Multiple Learning paths, including certification
  - Linkedin Learning (one month free trial, 279$ annual fee)
    - https://www.linkedin.com/learning/
- Use developer forums to ask and search for answers
  - Stackoverflow - https://stackoverflow.com/
  - https://softwareengineering.stackexchange.com/
  - https://dev.to/
  - https://www.experts-exchange.com/
  - https://www.quora.com/

Summary and take away

- Cloud is a platform of choice for Big Data and Data Analytics applications and tasks
- Hadoop is a standard de facto platform for Big Data Analytics
- All major CSP provide variety of Big Data Analytics services: AWS, Azure, GCP
- HDFS is a commonly recognised storage for Big Data and scalable data processing
- Data Lakes is a new model for Big Data and ELT (Extract – Load – Transfer) processes

Discussion Questions

- Go to **www.menti.com**

- What Big Data algorithm to use?

- Cloud is a solution for quick start and onboarding. What tasks are suitable for cloud?

- Storing data in cloud?

Group discussion

- Your organisation/project tasks related to Big Data and Data Analytics
- Mapping to AWS cloud resources