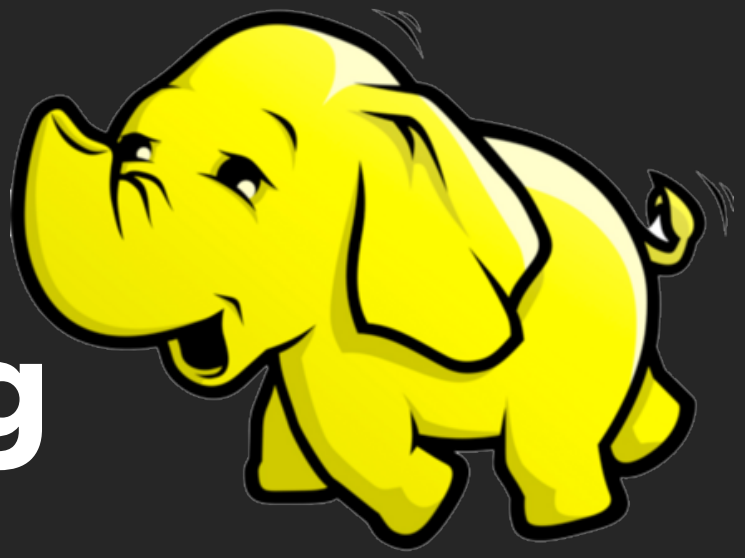
2nd Data Science Bootcamp, Lagos Nigeria. 12-15 October, 2017

Big Data Engineering with Hadoop



Johnson Iyilade, Ph.D.

CEO, Glomacs IT Solutions and Services Inc., Canada

Johnson.Iyilade@glomacssolutions.com

www.glomacssolutions.com

DSN Data Science Nigeria

2nd Data Science Bootcamp

GLOMACS

*"A data scientist is someone who is better at **statistics** than any **software engineer** and better at software engineering than any statistician."*

- Josh Wills

Table 1. The Ten Most Common Data Science Skills in Job Postings

The Ten Most Common Data Science Skills in Job Postings

Skill	Percentage of Job Listings
Python	72%
R	64%
SQL	51%
Hadoop	39%
Java	33%
SAS	30%
Spark	27%
Matlab	20%
Hive	17%
Tableau	14%

Source: Glassdoor Economic Research.

glassdoor

SOURCE:

Data Scientist Personas: What Skills Do They Have and How Much Do They Make? - September 21, 2017
Sample: **7,785 data science job**
https://www.glassdoor.com/research/data-scientist-personas/

Three Data Scientist Personas and What They Earn

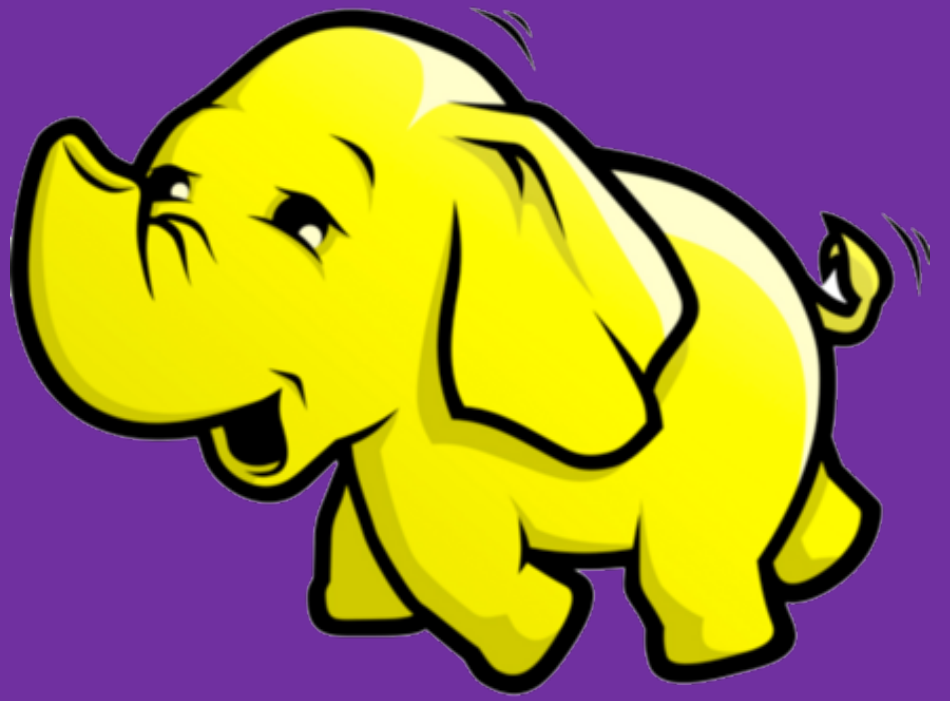
	Skills Likely to Have	Percentage of Data Science Jobs	Average Estimated Salary
Core Data Scientist	Python, R, SQL	**71%**	**$116,203**
Researcher	SAS, Matlab, Java, Hadoop, Python, R	**15%**	**$112,346**
Big Data Specialist	Spark, Hive, Hadoop, Java, Python	**14%**	**$121,246**

Source: Glassdoor Economic Research.

glassdoor

Data Scientist Personas: What Skills Do They Have and How Much Do They Make? - September 21, 2017
Sample: **7,785 data science job**
https://www.glassdoor.com/research/data-scientist-personas/

What is Big Data?

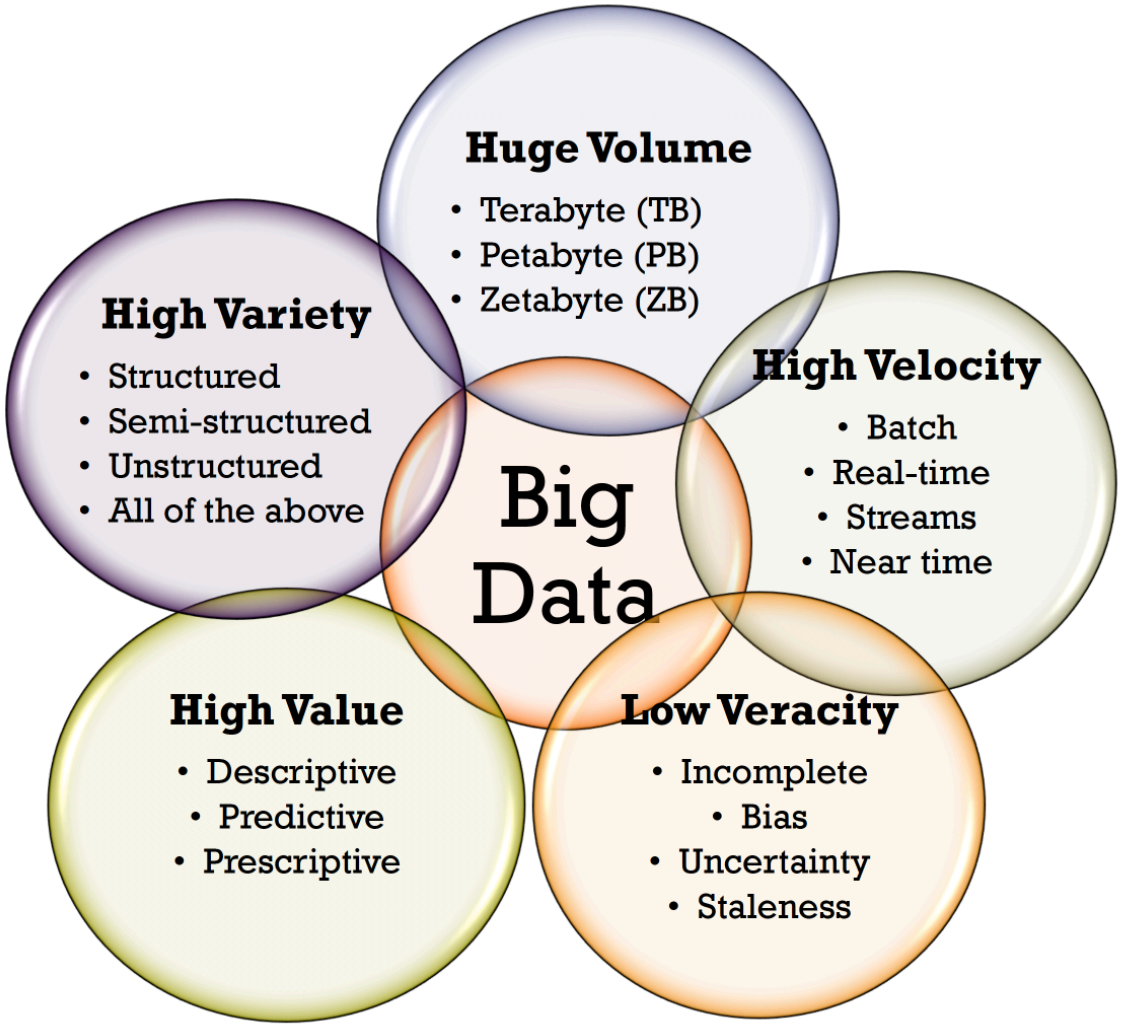


GLOMACS

Defining Big Data?

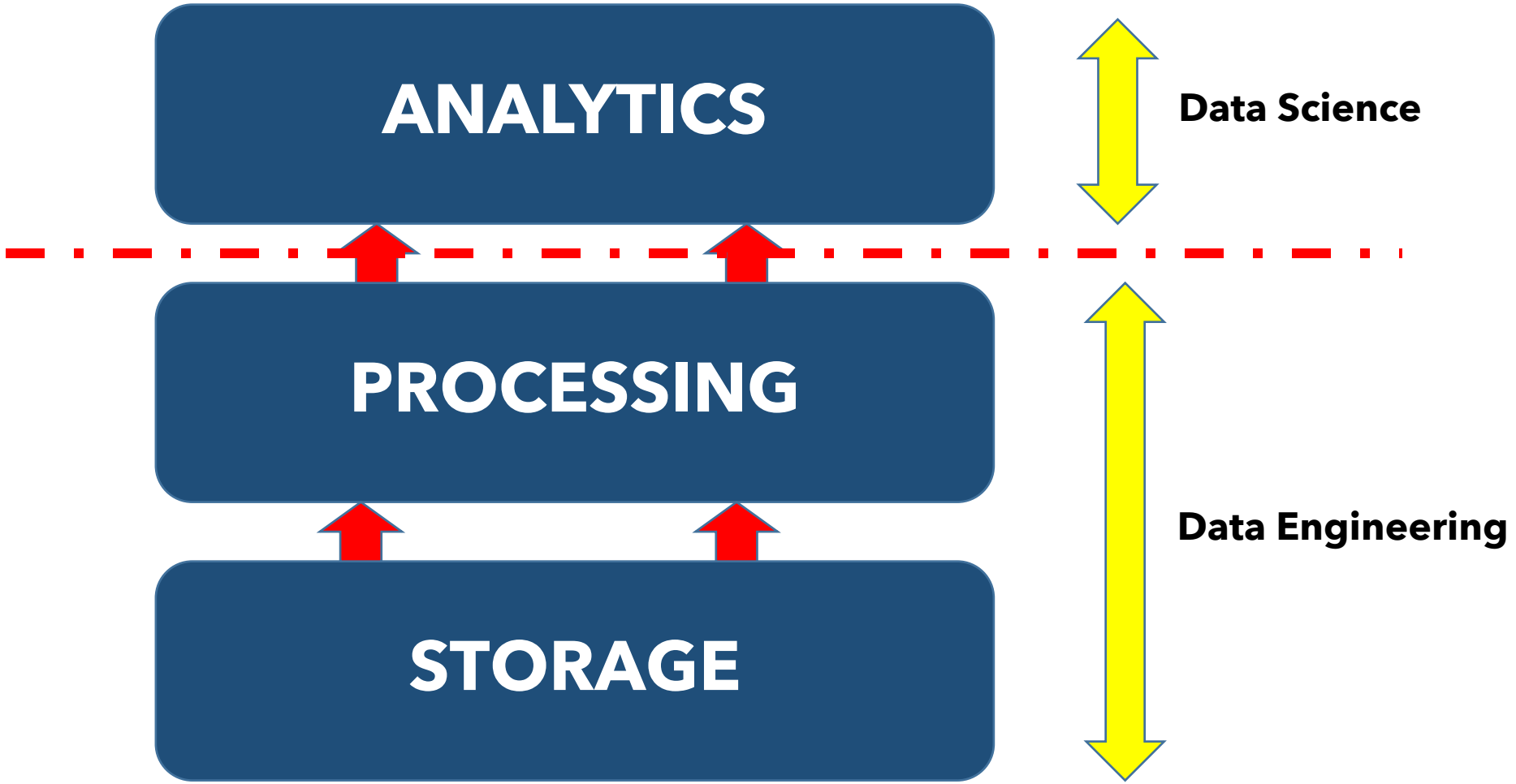
Extremely large and varied datasets that may be analyzed to reveal patterns, trends, and associations; they're often too expensive to store, process and analyze with traditional storage and computing methods.

5V's Big Data



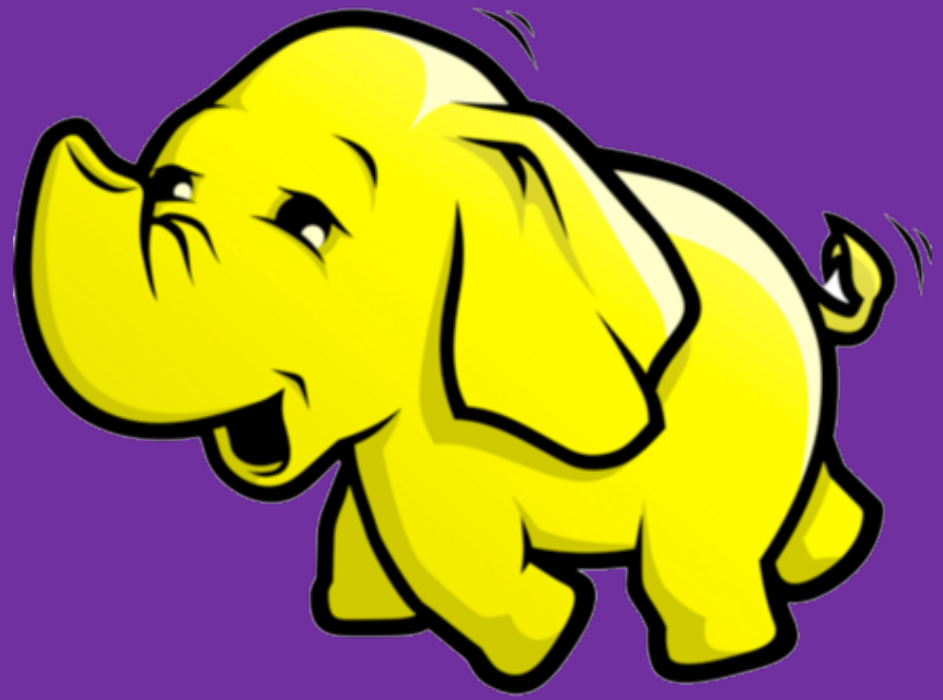
4 of these 5V's
(Volume, Velocity, Veracity, Variety)
are mostly
Data Engineering
challenges

3-Tier Big Data Architecture



GLOMACS

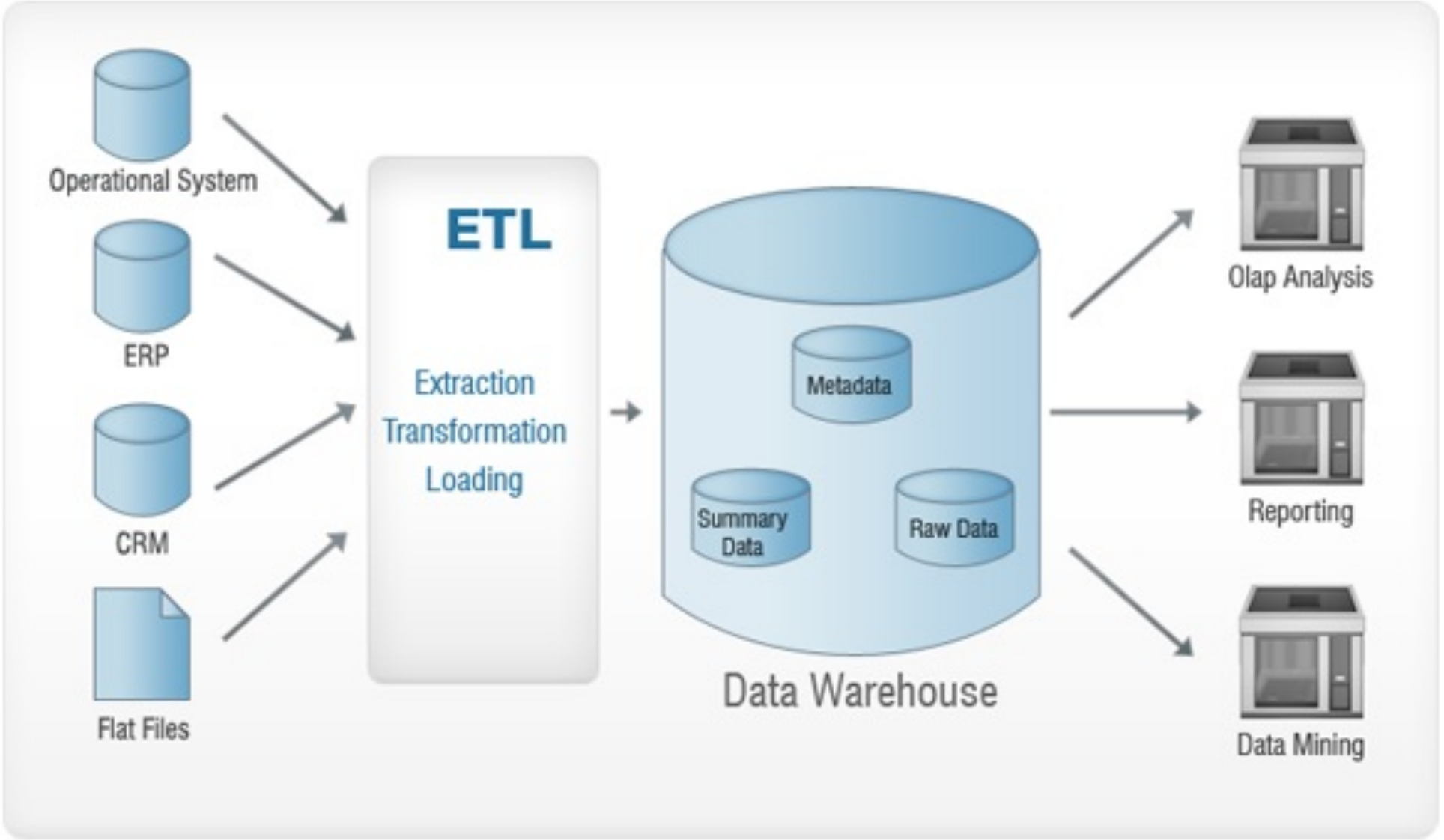
Big Data Disruption in the Enterprise



GLOMACS

Businesses have always recognize and collect data as an integral part of business for understanding patterns and insights on customers, competitors and markets

Data Warehouse (DW) and Business Intelligence (BI) had Traditionally become the means for meeting businesses' needs



* *EDW promoted a* ***single, central version of the truth*** *for an organization*

* *A repository to* ***gather and integrate data*** *to quickly and easily create reports.*

* *Simplify data* ***access*** *and* ***reporting***

* *Combining data from many sources to* ***answer questions business may have***

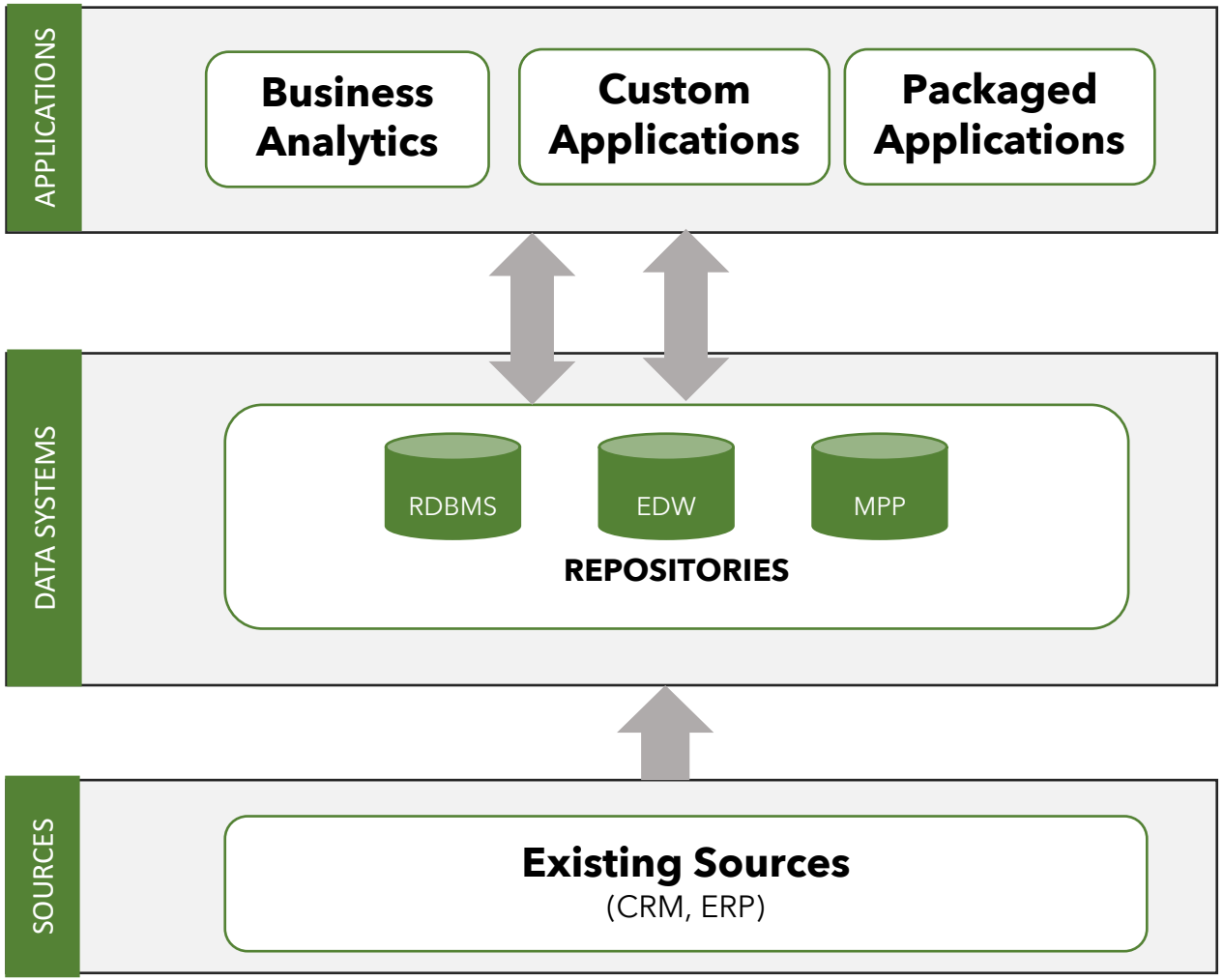
Why do businesses need something more

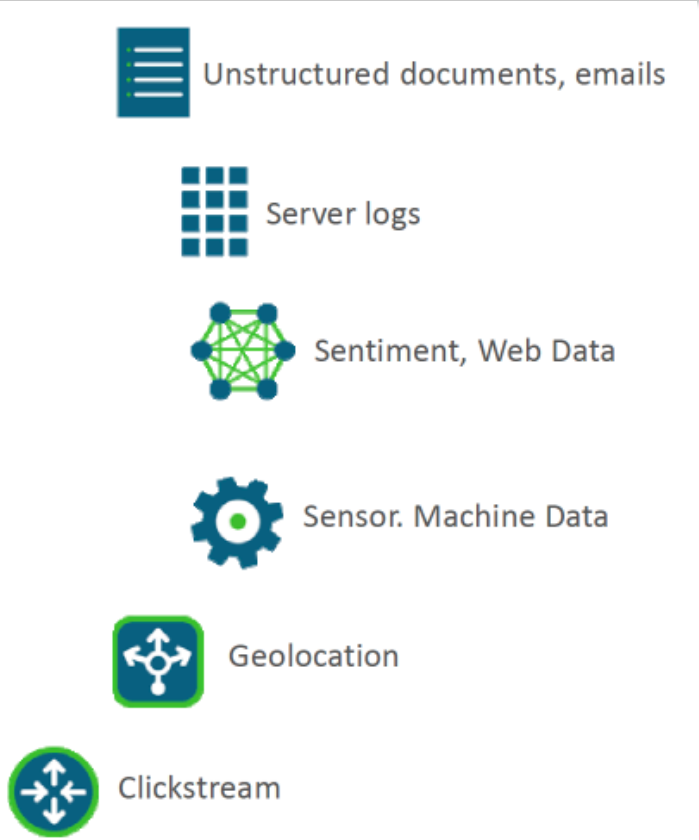
than what traditional EDW and BI reporting offered?

1

It's impossible to anticipate every question a business might ask and every use-case or report they might need at the point of storing the data

2 **Traditional solutions** under pressure from flood of new data types and sources





In traditional Data Warehouse
Organizations will ignore these data because they are either too voluminous or in a format that is difficult to manipulate and store

3 **New Types of Data** provides richer and broader insights about customer (Customer 360)

Sentiment
Understand how your customers feel about your brand and products – right now

Clickstream
Capture and analyze website visitors' data trails and optimize your website

Sensors
Discover patterns in data streaming automatically from remote sensors and machines

Geographic
Analyze location-based data to manage operations where they occur

Server Logs
Research logs to diagnose process failures and prevent security breaches

Unstructured
Understand patterns in files across millions of web pages, emails, and documents

Most enterprise data today are coming from **"systems of engagement"** such as website clickstream, mobile, social media or from connected devices (IoTs) rather than traditional **"system of storage"** (CRM, ERP, etc)

Evolution From **System-Centric** to **User-Centric** Connected Data Architectures

IDMS

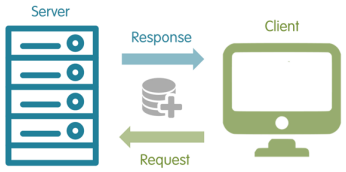
Relational Database

Connected Data Lake
- Collect All Data:
Data in Motion and
Data at rest
- Cloud/Data Center
- Powered by Open Source Platform

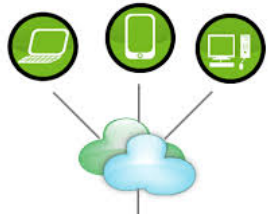
SYSTEM-CENTRIC
SYSTEM OF STORAGE

USER-CENTRIC
SYSTEM OF ENGAGEMENT

Mainframe



Client-Server



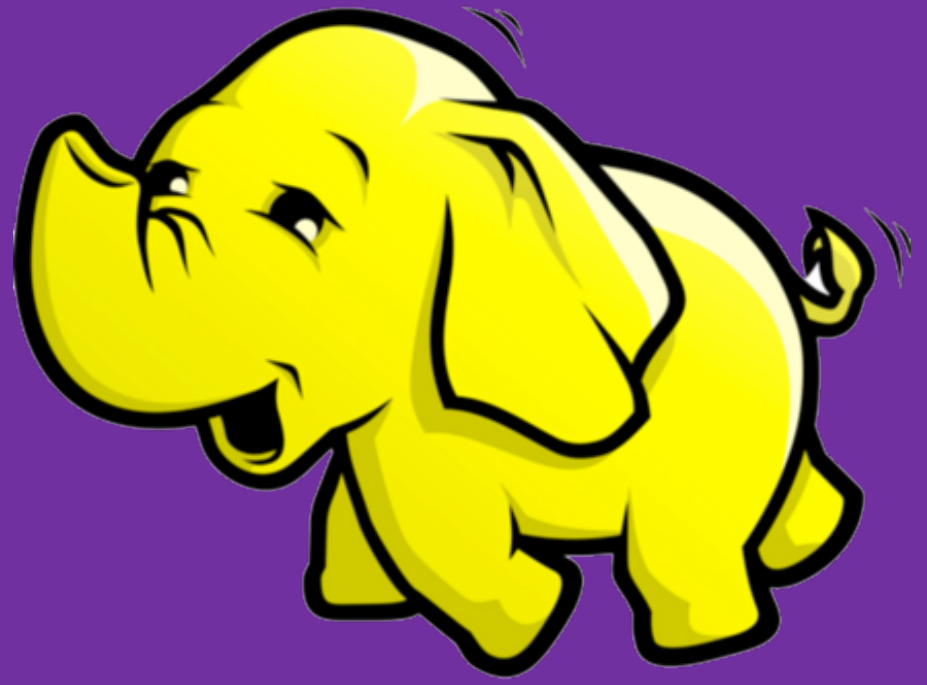
Web/SaaS

Modern Data Apps

New Data Use Cases
*Connected cars
*Factory Automation
*Predictive Analytics
*AI

In recent years, the limitations of traditional Enterprise data warehouse has spawned a new set of architecture, platform, technologies, tools and techniques for storing and processing data

Apache **Hadoop** & Modern **Enterprise** **Data Lake** to the rescue…



GLOMACS

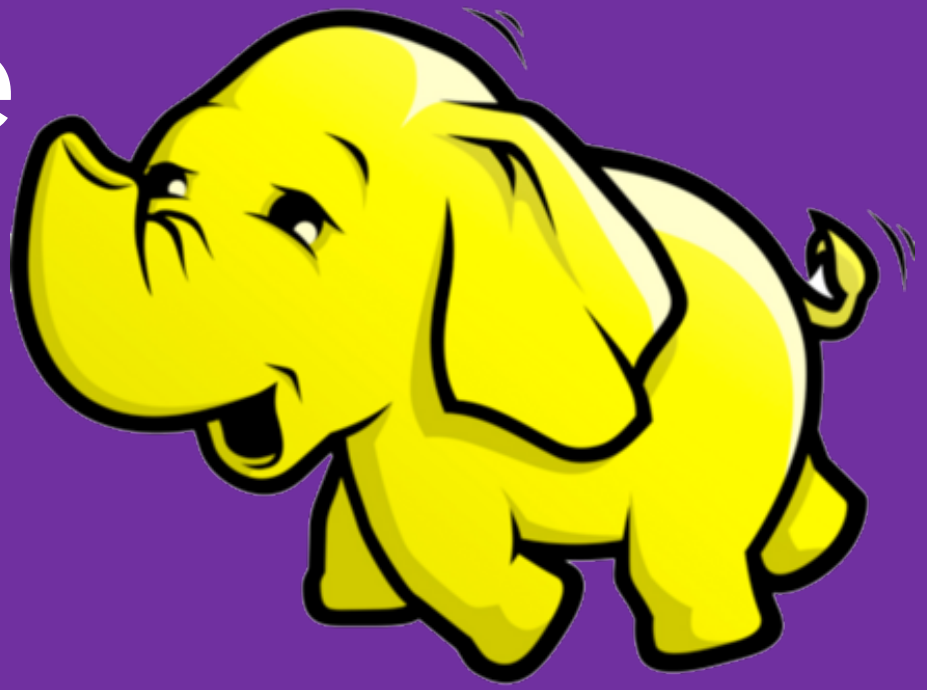
Hadoop is…

…a **FRAMEWORK** of OPEN SOURCE **tools, libraries** or **methodologies** for **storage** and **processing** extremely large data sets in a distributed computing environment

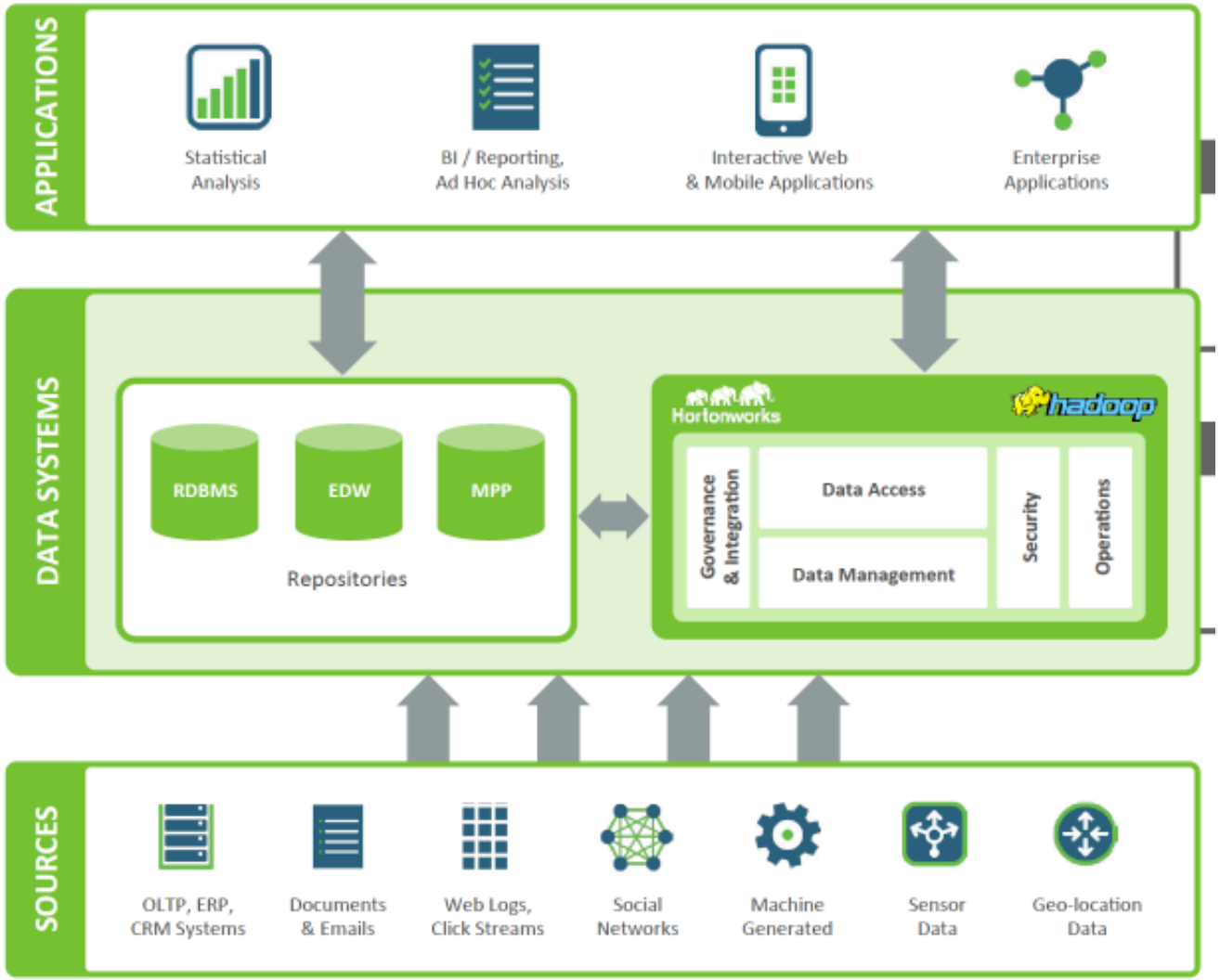
Hadoop CHARACTERISTICS

- **Open source** (Apache License)
- Can **handle large** unstructured **data sets**
- **Scalable** from single server to thousands of machines
- Runs on **commodity hardware** or the **cloud**
- Application-level **fault tolerance**
- Multiple **tools and libraries** integrated

Modern enterprise data architecture with Hadoop



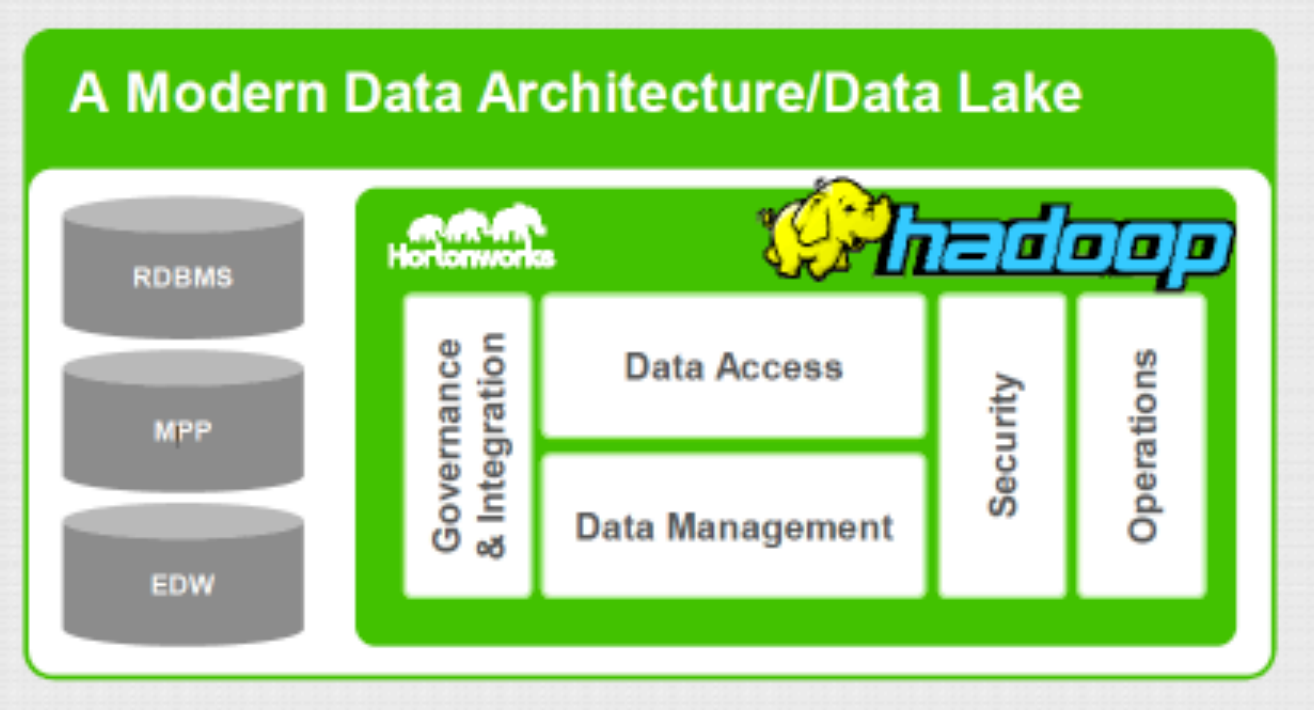
Modern data architecture
Apache Hadoop + existing data systems



HADOOP AND DATA LAKE…

- Complimentary to and co-exist with existing systems (EDW and BI). Organization can continue to leverage their existing investment in EDW and BI while collecting new data they've been throwing away.
- Low Cost approach to data storage and processing
- Scale Out
- Wider Scope

Hadoop has moved enterprises toward the vision of a '**Data Lake'**



(c)Hortonworks

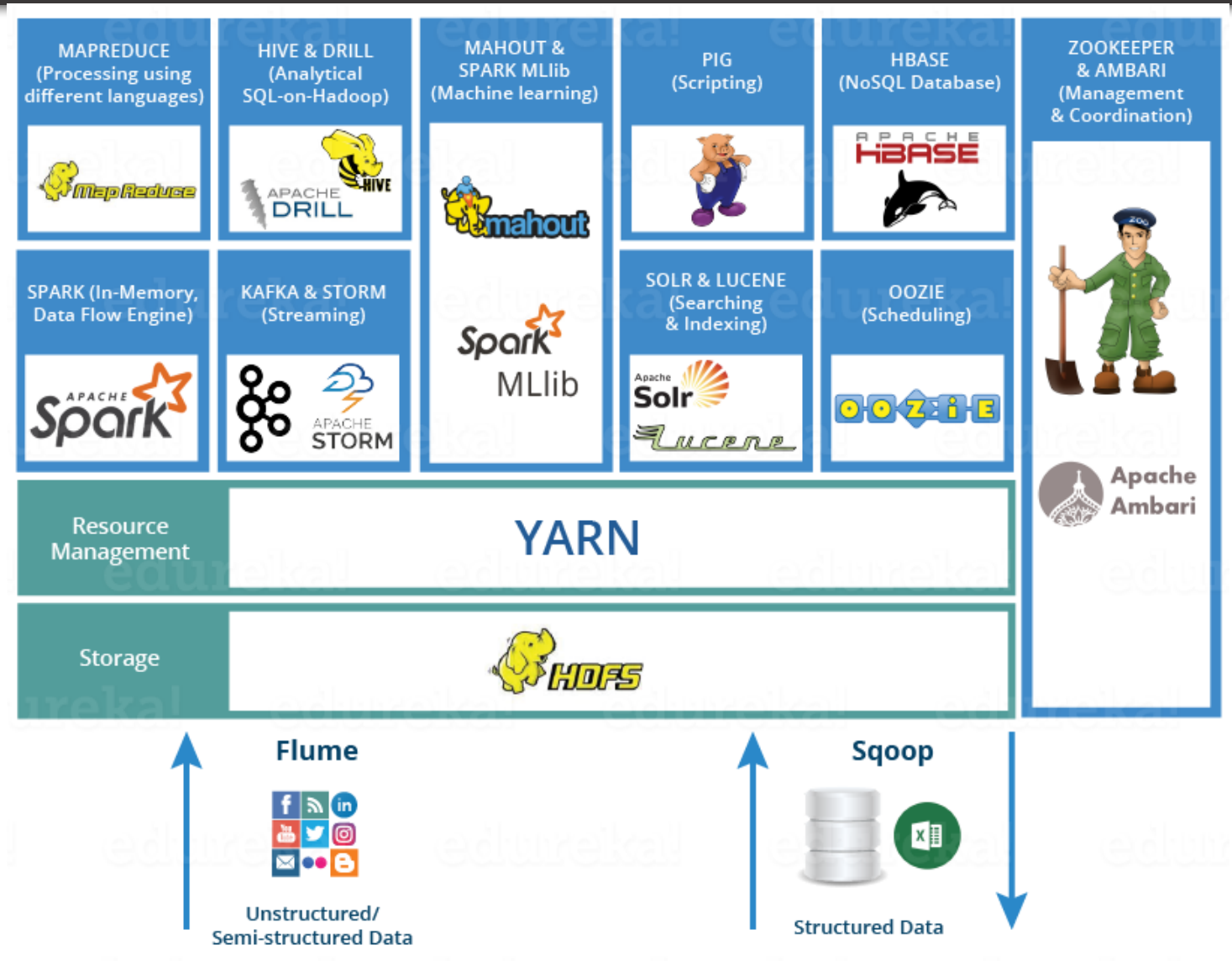
ENTERPRISE DATA LAKE...*is an arsenal to* ***store vast amounts of raw data for future use***.

- Data flows from the streams (the source systems) to the lake
- All data is loaded from source systems. No data is turned away.
- Data is stored in an untransformed or nearly untransformed state.
- Data is transformed and schema is applied to fulfill the needs of analysis.

Differences between a **data lake** and a **data warehouse**

1. Data Lakes Collect and Retain All Data
2. Data Lakes Support All Types of Data
3. Data Lakes Support All Users
4. Data Lakes Provide Faster Insights

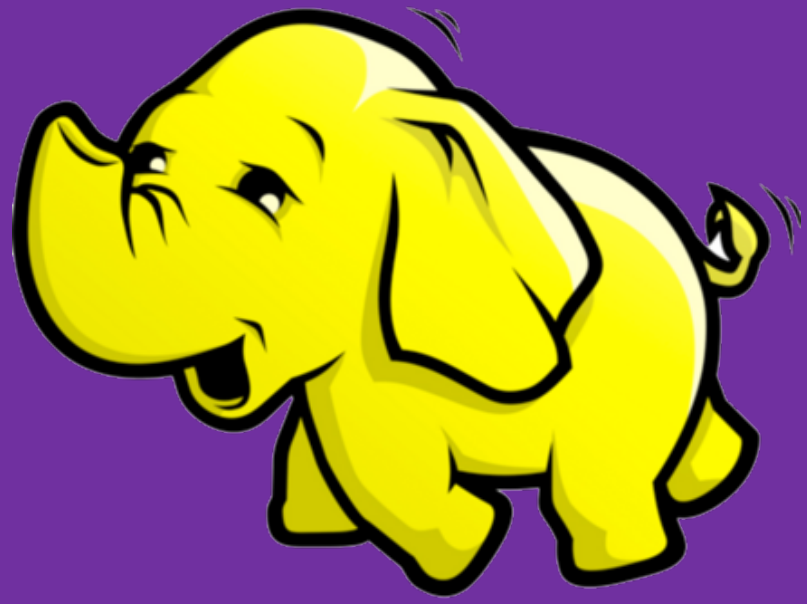
Hadoop Ecosystem **Platform & Tools**



Tools & Usage

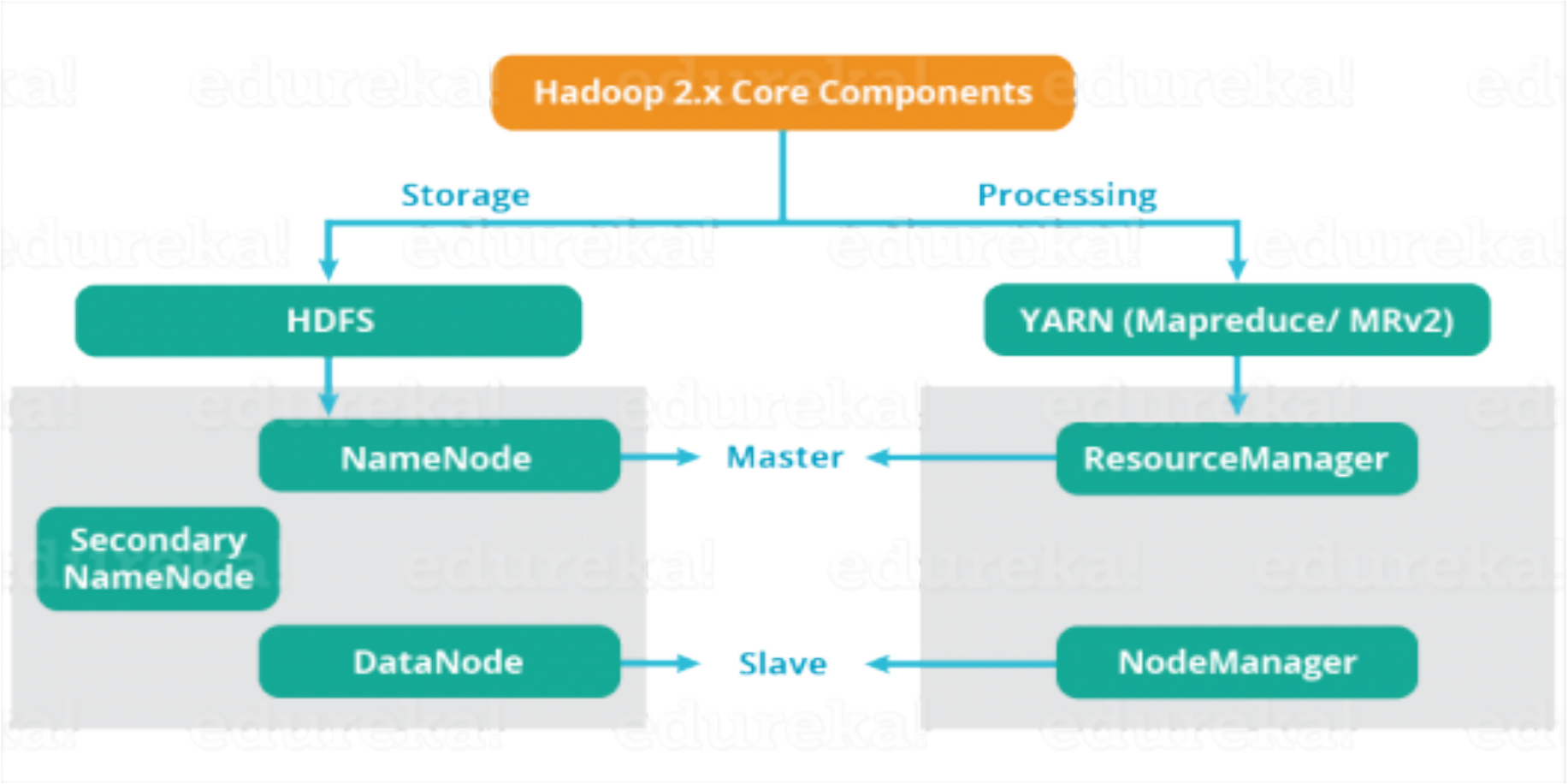
TOOL	USAGE
Ambari	Deployment, configuration and monitoring
Flume	Collection and Import of log and event data
HBase	Column-oriented database scaling to billions of rows
HCatalog	Schema and data type sharing over Pig, Hive, and MapReduce
HDFS	Distributed redundant file system for Hadoop
Hive	Data Warehouse with SQL-like access
Mahout	Library of Machine Learning and DM Algorithma
MapReduce	Parallel Computation on Server Clusters
Pig	High-level programming for Hadoop Computations
Oozie	Orchestration and workflow management
Sqoop	Imports data from relation databases
Zookeepers	Configuration Management and Coordination
Spark	Framework for Real-time Analytics

Storage & Processing in Hadoop



GLOMACS

Hadoop
Core Components

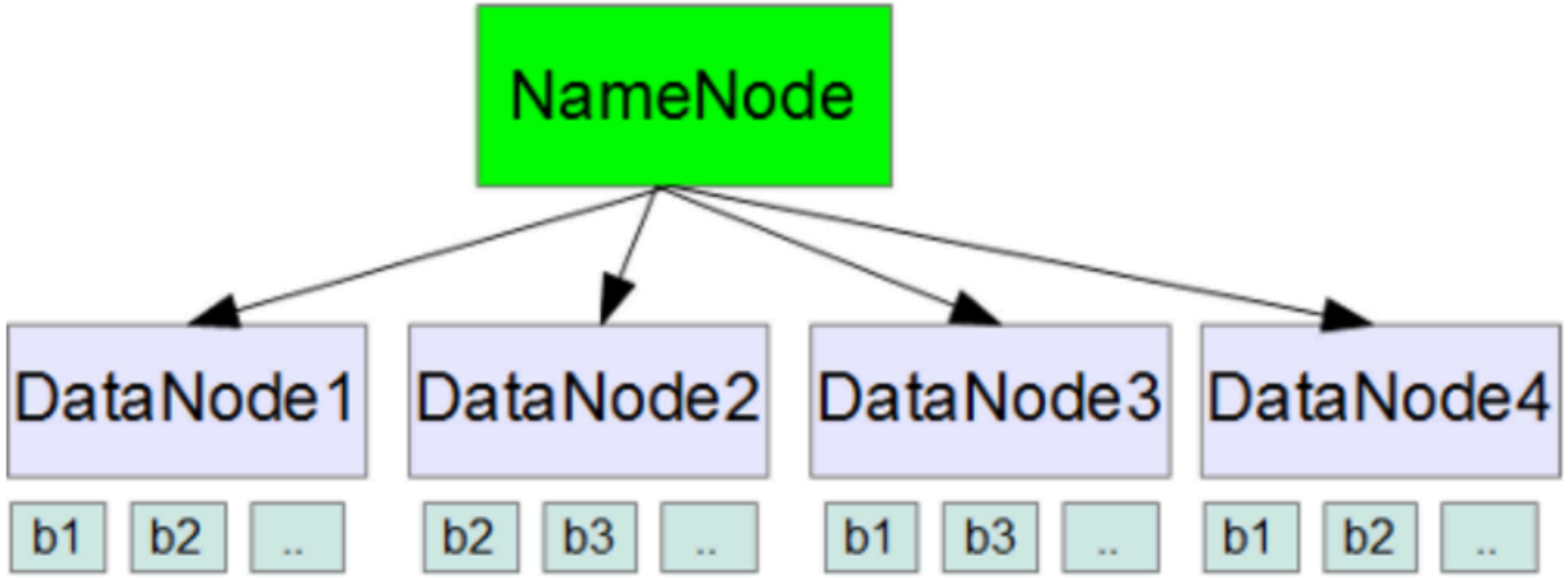


HDFS

HADOOP DISTRIBUTED FILE SYSTEM

- HDFS is a core component or backbone of Hadoop ecosystem.
- HDFS makes it possible to store different types of large data sets (i.e. structured, unstructured and semi structured data).
- HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit.
- Helps in storing data across various nodes and maintaining the log file about the stored data (metadata)
- Has two core components, i.e. **NameNode and DataNode**

HDFS HADOOP DISTRIBUTED FILE SYSTEM

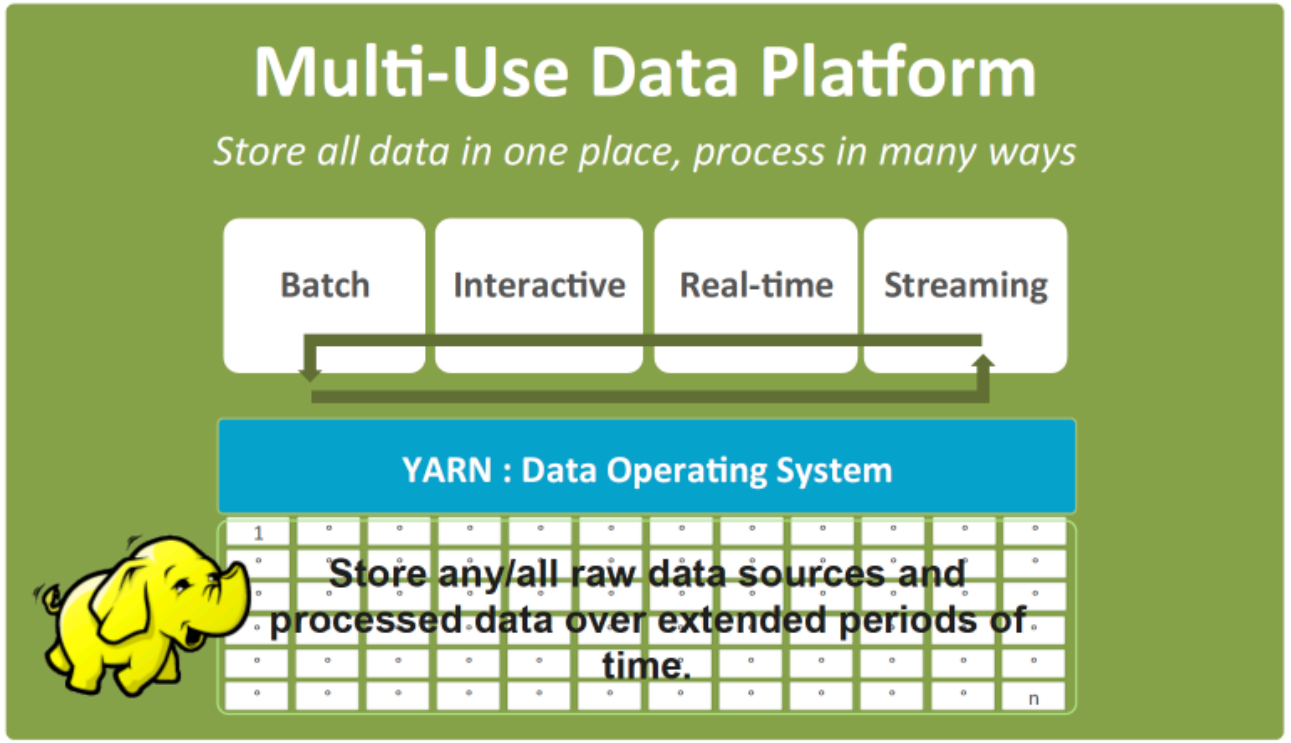


YARN

YET ANOTHER RESOURCE NEGOTIATOR

- YARN is like the “brain” of the Hadoop Ecosystem
- Often referred to as the DATA OPERATING SYSTEM
- Performs all processing activities by allocating resources and scheduling tasks
- YARN allows multiple data processing engines such as interactive SQL, real-time streaming, data science and batch processing to handle data stored in a single platform
- The components of ***YARN*** are **ResourceManager** and **NodeManager**.

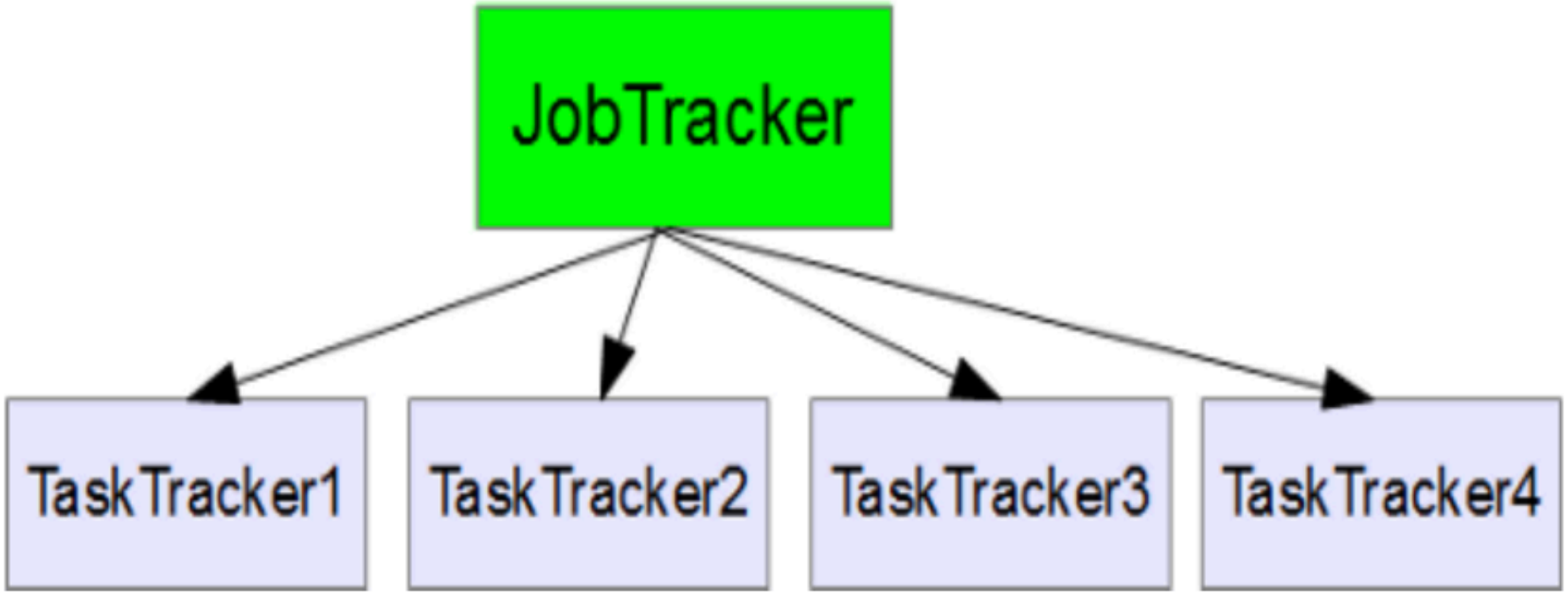
YARN transforms Hadoop Architecture



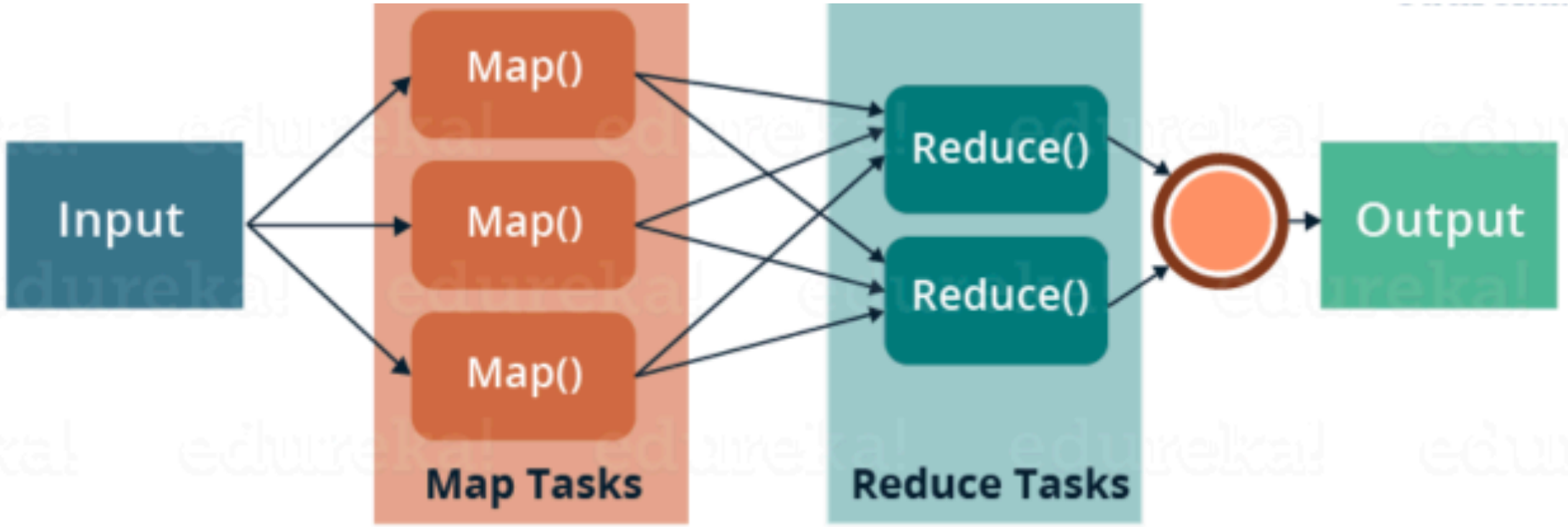
MapReduce

- A programming framework for distributed and parallel processing on large data sets
- Consists of two distinct tasks – **Map** and **Reduce**.
- During **Map** job, a block of data is read and processed to produce key-value pairs as intermediate outputs. The output of a Mapper or map job (key-value pairs) is input to the Reducer.
- The **Reducer** receives the key-value pair from multiple map jobs. The reducer aggregates those intermediate data tuples into a smaller set of tuples or key-value pairs which is the final output.

MapReduce COMPONENTS

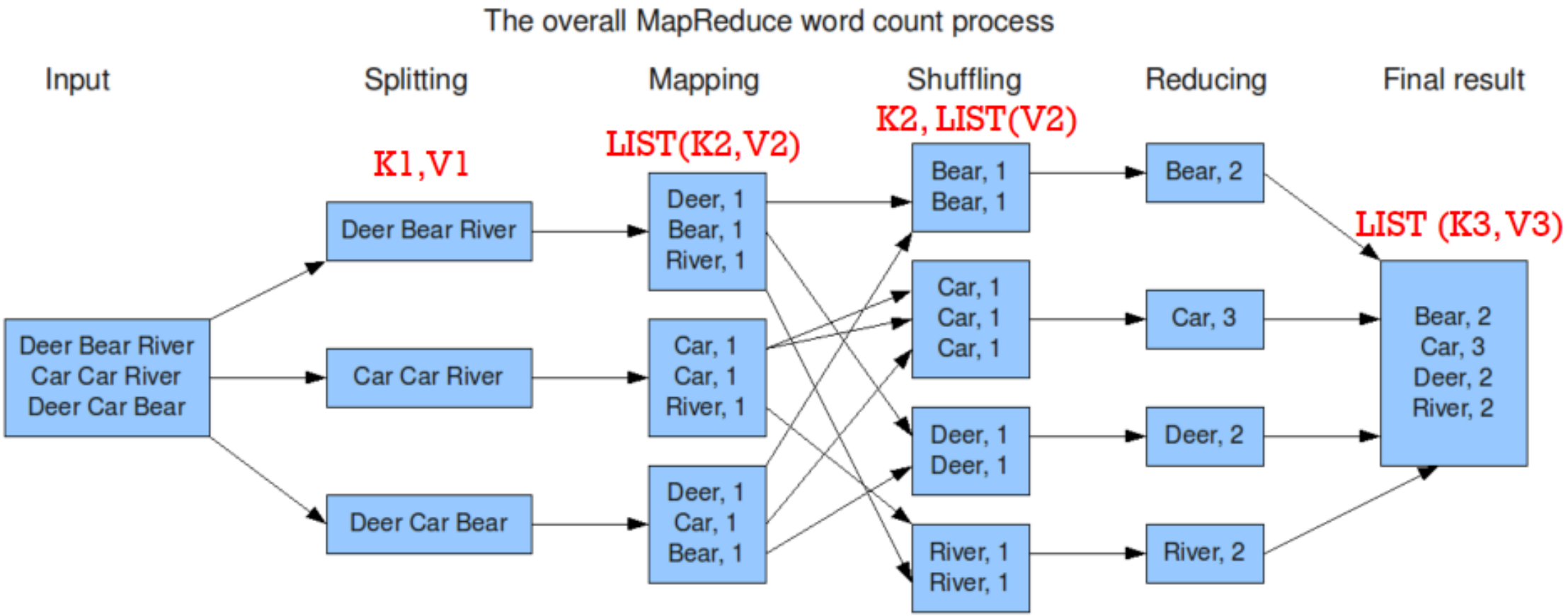


MapReduce PROGRAMMING **MODEL**

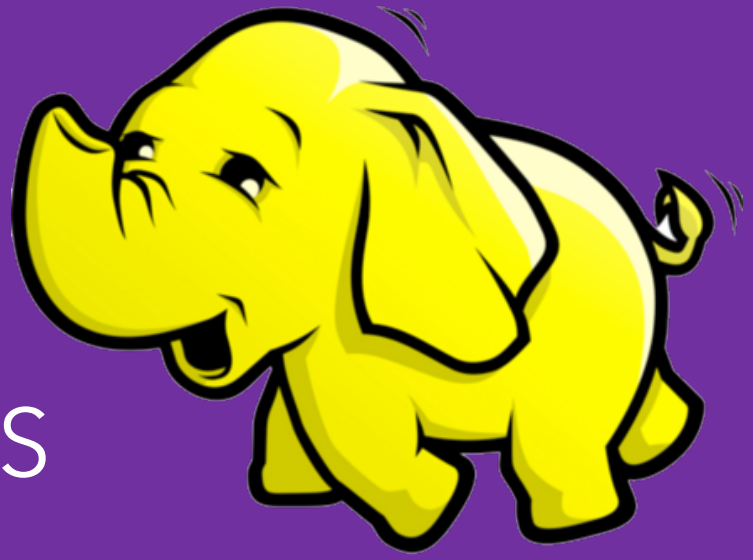


MapReduce

EXAMPLE: **WORD COUNT**



Hadoop DISTRIBUTIONS



GLOMACS

Hadoop PLATFORM **DISTRIBUTIONS**

Cloudera

- Cloudera Inc. was founded by a group of big data geniuses from Google, Oracle, and Yahoo in 2008
- First company to develop and distribute Apache Hadoop-based software and still has the largest user-base and clients
- Provides proprietary Cloudera Management Suite (CMS) to automate the installation and provide other services to enhance convenience of users

Hortonworks

- Founded in 2011, and has quickly emerged as one of the leading vendors of Hadoop.
- Provides open source platform based on Apache Hadoop for storing and managing big data
- The only commercial vendor to distribute complete open source Apache Hadoop without additional proprietary software
- They are behind most of Hadoop's recent innovation including YARN

MapR

- Another Hadoop distribution but replaces HDFS component and instead uses its own proprietary file system called MapRFS.
- MapRFS helps incorporate enterprise-grade features into Hadoop, enabling more efficient management of data, relability and ease of use
- More production ready that the other two
- Recently, MapR is offering Hadoop as a default component of Ubuntu operating system
- Up to it's M3 edition, MapR is free, but free version lacks some of its proprietary features.

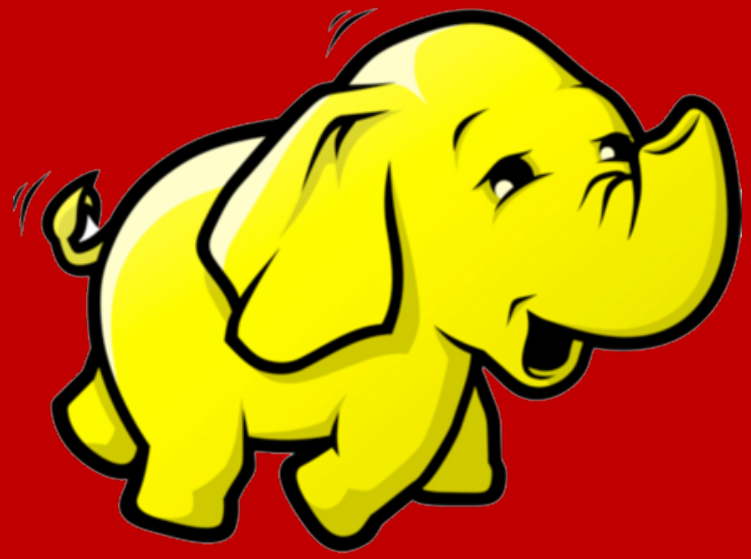
Cloudera vs Hortonworks: Similarities

- Both offer enterprise-ready Hadoop distributions
- Both have established communities
- Both distributions have master-slave architecture.
- Both have a shared-nothing computing framework
- Both support MapReduce as well as YARN

Cloudera vs Hortonworks: Differences

- Cloudera intends to become *enterprise data hub* thereby diminishing the need for EDW. Hortonworks remain a hadoop distributor
- While Cloudera CDH can be run on windows server, HDP is available as a native component on the windows server through HDInsight service
- Cloudera has a proprietary management software Cloudera Manager, SQL query handling interface Impala, as well as Cloudera Search for easy and real-time access of products. Hortonworks has none
- Cloudera has a commercial license, while Hortonworks has open source license
- Cloudera has a free 60-day trial, Hortonworks is completely free.

Thank You!



www.glomacssolutions.com
Johnson.Iyilade@gmail.com

GLOMACS