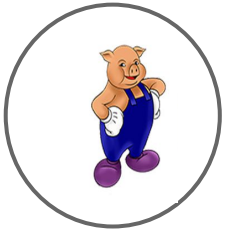
Hadoop Ecosystem

Hadoop Ecosystem

Kafka

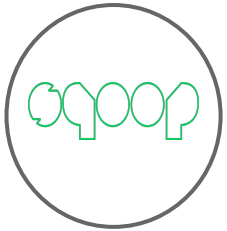


Pig

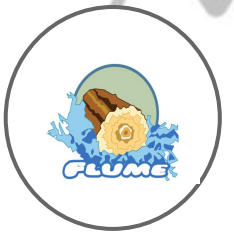
Hive

Spark

HBase



Sqoop

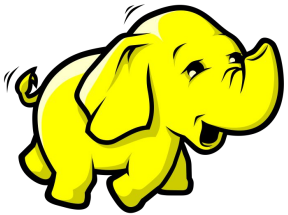


Flume

Hadoop

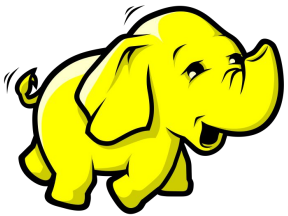
Impala

Zookeeper



Hadoop Distributed File System (HDFS)

- Storage Layer for Hadoop
- Distributed Storage
- Highly Scalable
- Runs on Commodity Hardware



MapReduce

- Main processing engine of Hadoop
- Consists of two parts: Map and Reduce tasks
- Fault tolerant
- Parallel Computation

NoSQL Database

Stores data in HDFS

Random read/write

Real-time read/write

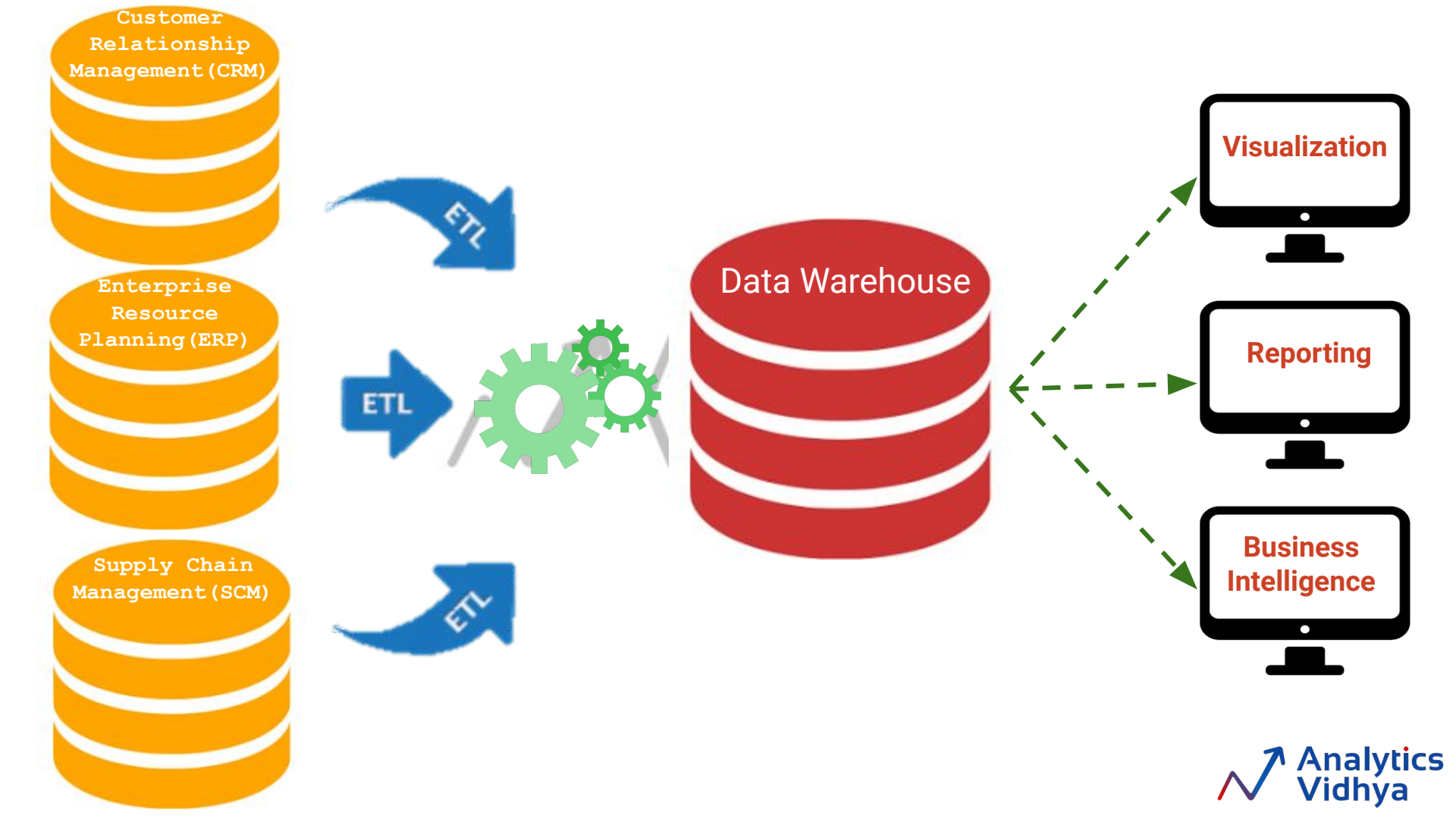
Apache Pig

Abstraction over Map-Reduce

Analyze large dataset

Uses Pig Latin

- Distributed data warehouse system
- Supports Hive Query Language (HQL)
- Executes queries using map-reduce.
- Used for Analytical Jobs


Customer Relationship Management (CRM)
Enterprise Resource Planning (ERP)
Supply Chain Management (SCM)
ETL
ETL
ETL
Data Warehouse
Visualization
Reporting
Business Intelligence
Analytics Vidhya

APACHE
ZooKeeper™

- Managing the cluster
- Naming Service
- Distributed coordination service

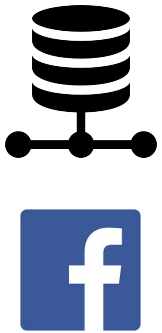
Handles real-time streaming data

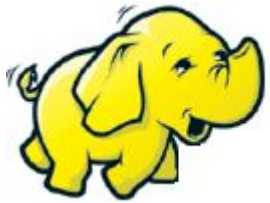
Ingests streaming data from various sources

Streaming data to various applications

Collect data from multiple data sources in Hadoop

Can collect data in real-time as well as in batch mode

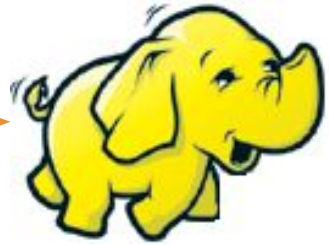




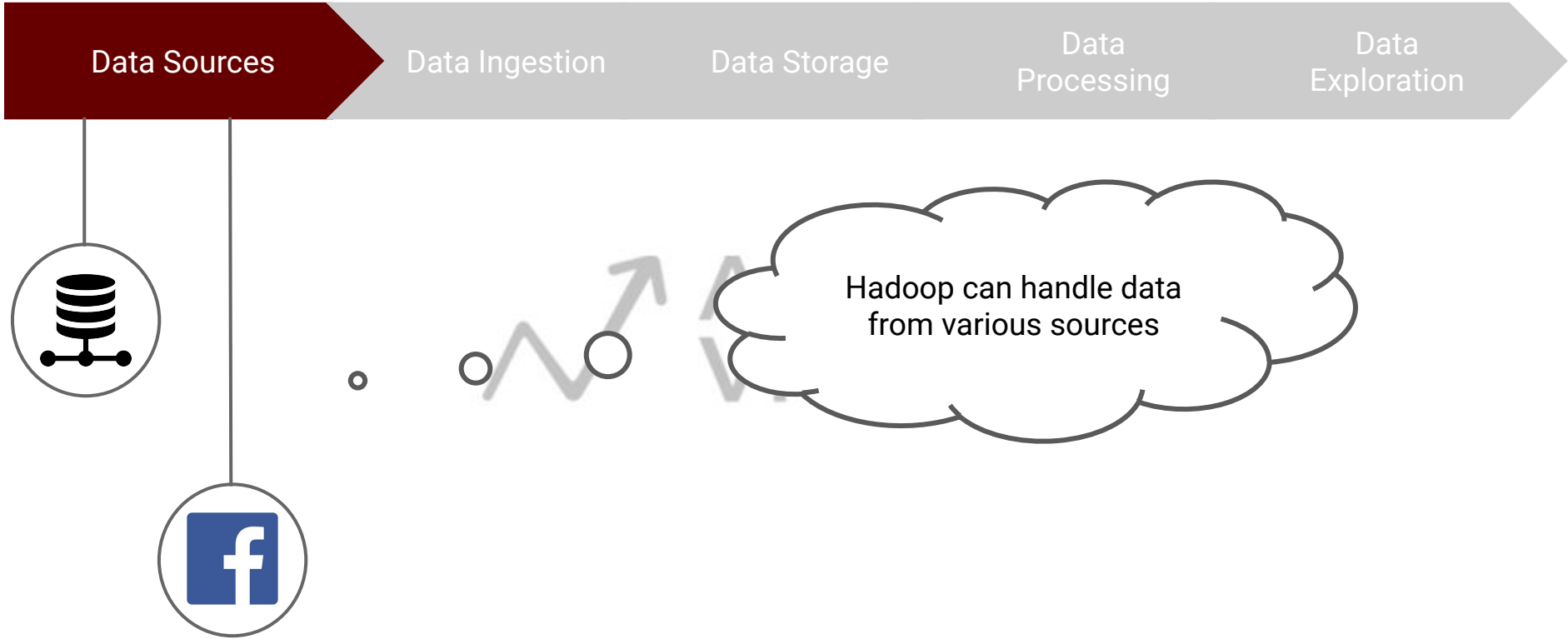
Robust and fault-tolerant

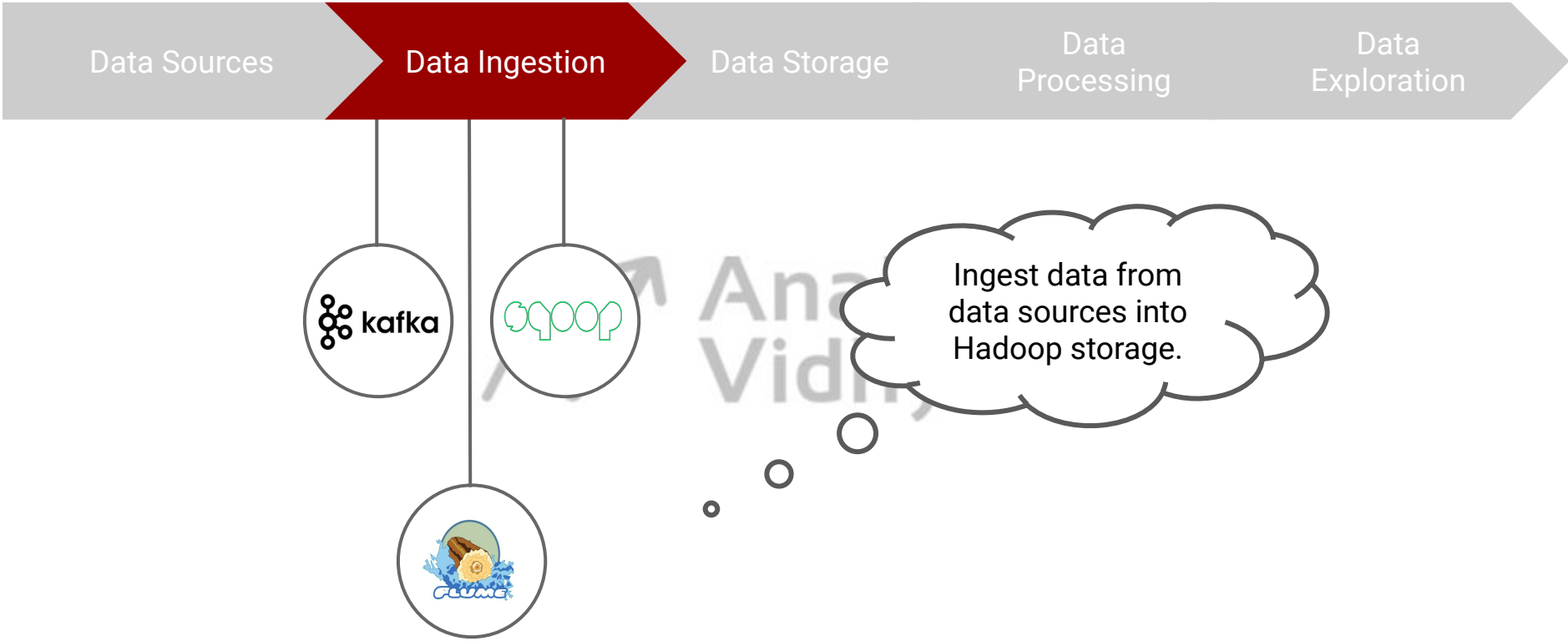
Parameter	Apache Kafka	Flume
Push/ Pull Model	Kafka is works as a pull model.	Flume is works as a push model.
Recovery	Highly available and resilient to node failures.	In case of Flume-agent failure, you will lose events in the channel.
Flexibility	Kafka is a general purpose publish-subscribe model messaging system	It is specially designed for Hadoop

Can import data from RDBMS into Hadoop

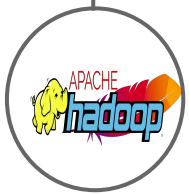


Can export data from Hadoop into RDBMS

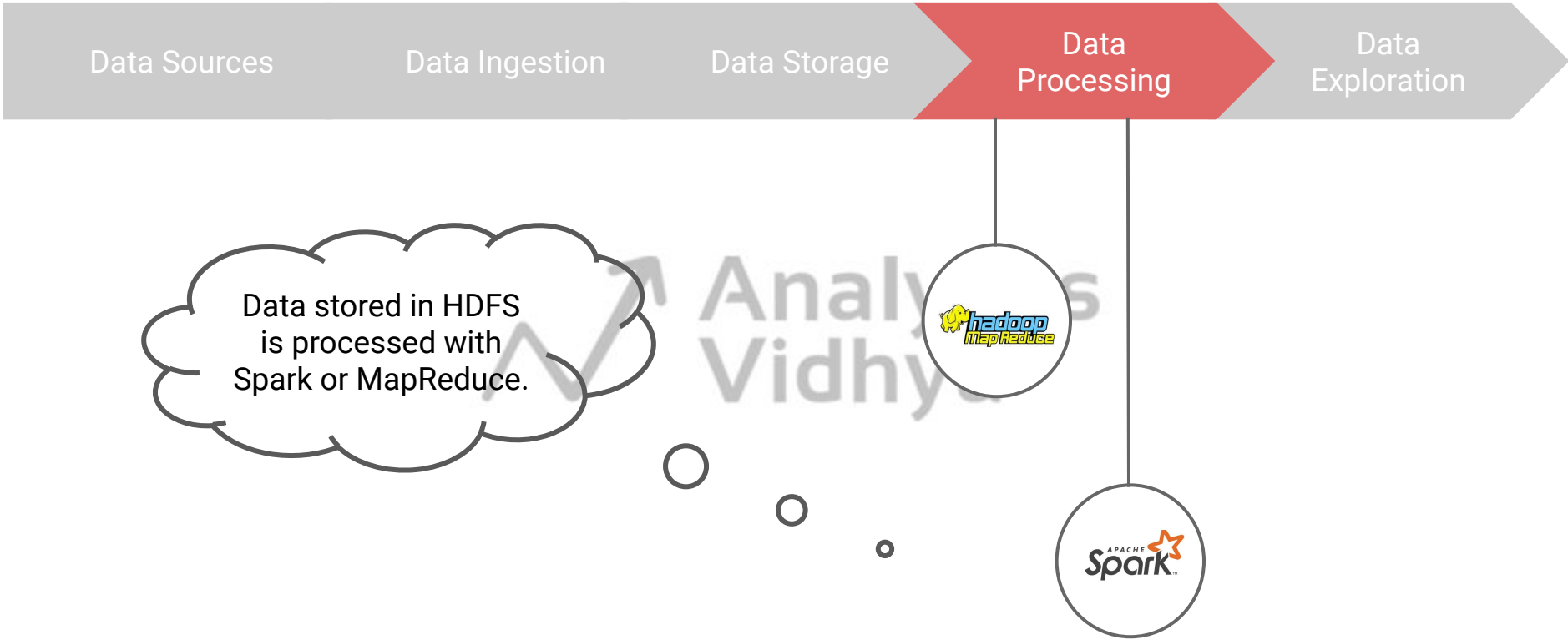




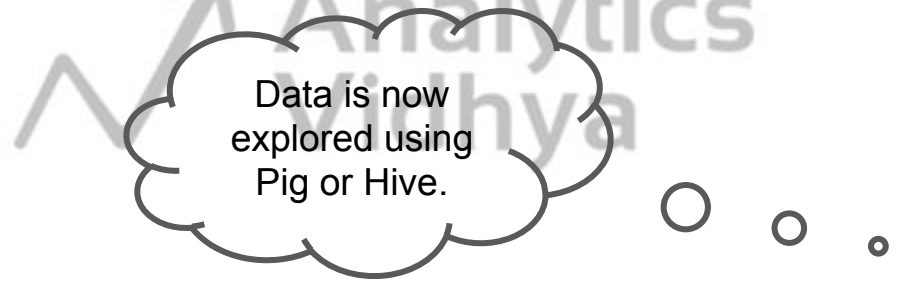
Data Sources | Data Ingestion | **Data Storage** | Data Processing | Data Exploration

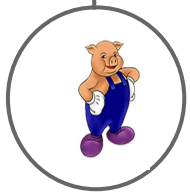


HDFS is the distributed storage component of the Hadoop ecosystem.



Data Sources | Data Ingestion | Data Storage | Data Processing | Data Exploration





Hadoop Characteristics

Reliable

* Stores multiple copies of data on different nodes

* Resistant to hardware failures

Flexible

* Can store lots of data

* Can store structured or unstructured data

Scalable

* Can add lots of nodes to the cluster

* Can scale nodes vertically as well

Economical

* Nodes are commodity hardwares

Thank You!!

Analytics
Vidhya