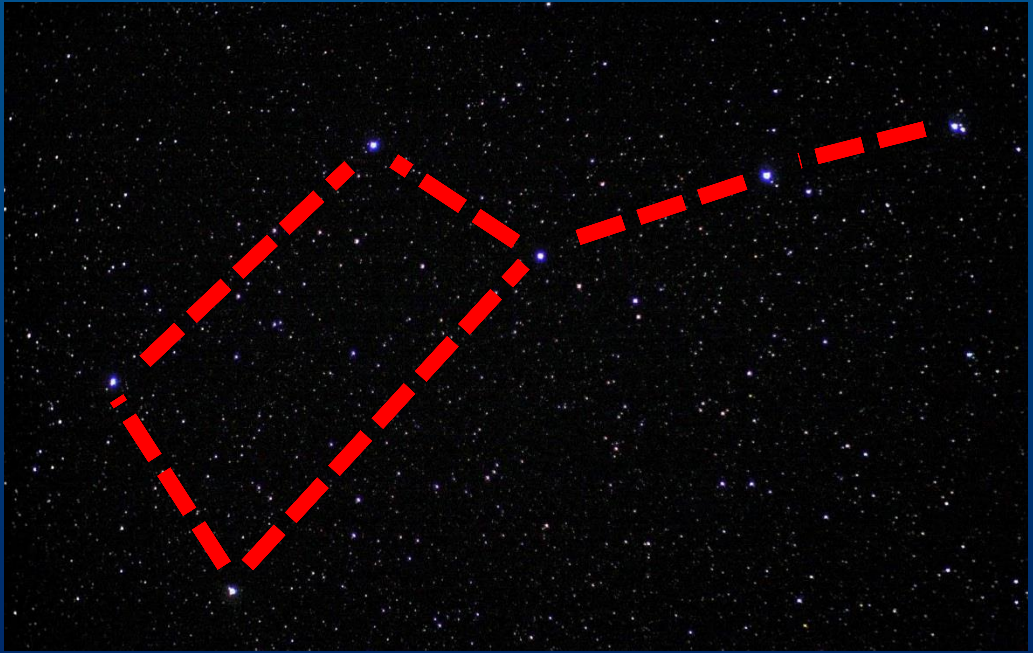
Big Data Analytics

Welcome and Introduction

The big picture: you have lots of data …

… and you want to see meaningful things.



Meaningful things give insight!

Analysis

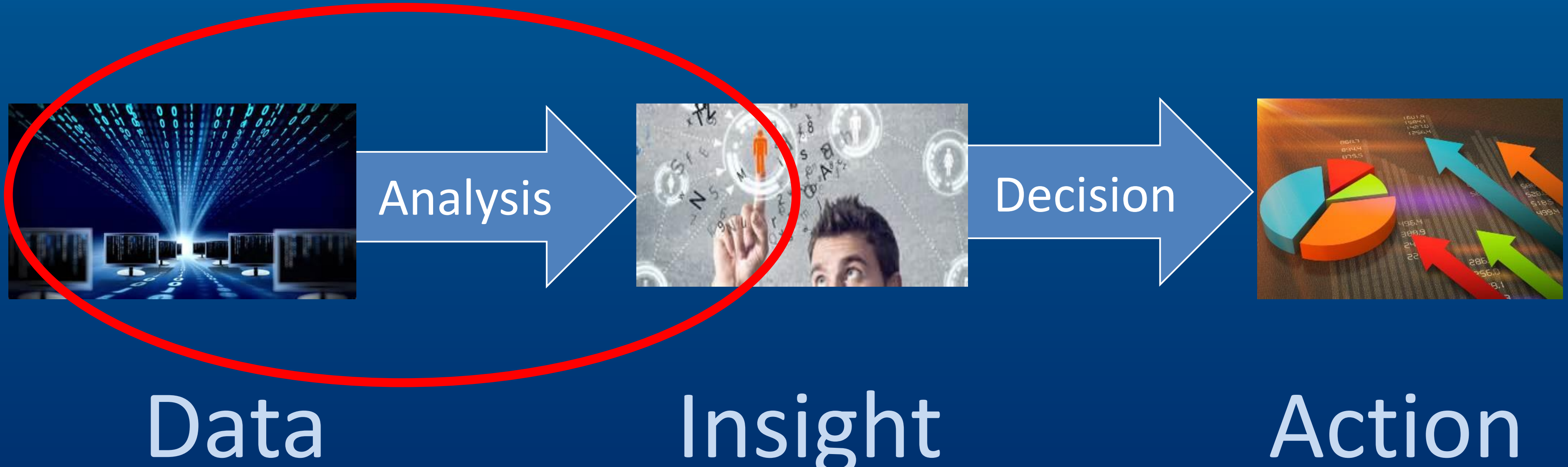
Decision

Data

Insight

Action

This course and next course



Analysis

Decision

Data

Insight

Action

Analysis (in a nutshell)

- Search for:

  Interesting observations

  Summaries

  etc …

The Analyst Needs:

The Analyst Needs:

- Easy access to data

The Analyst Needs:

- Easy access to data
- Good data functions

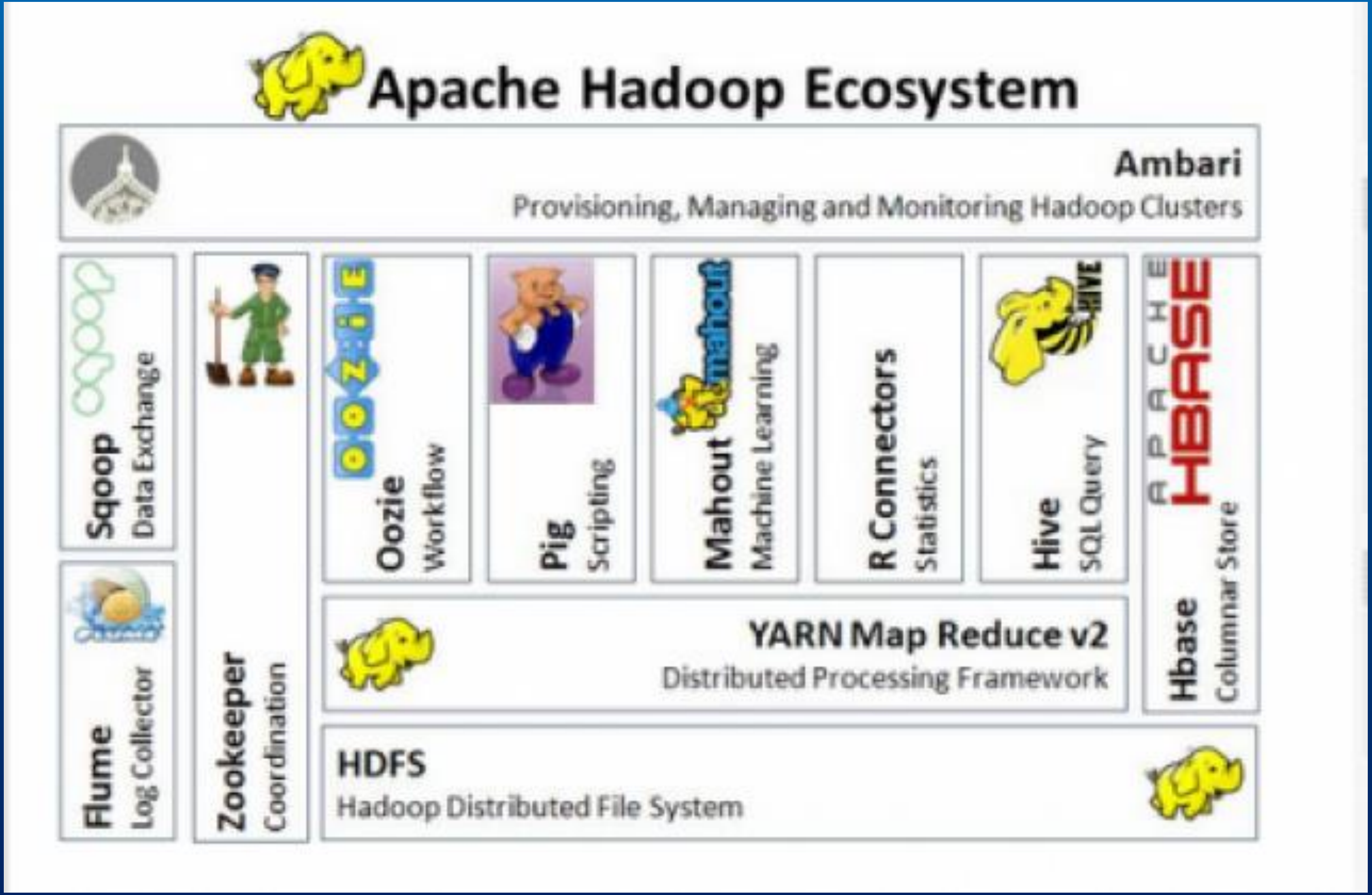
The Analyst Needs:

- Easy access to data
- Good data functions
- Query/Exploratory support

The Analyst Needs:

- Easy access to data
- Good data functions
- Query/Exploratory support

We Need Hadoop Tools!



Apache Hadoop Ecosystem

Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

Sqoop
Data Exchange

Flume
Log Collector

Zookeeper
Coordination

Oozie
Workflow

Pig
Scripting

Mahout
Machine Learning

R Connectors
Statistics

Hive
SQL Query

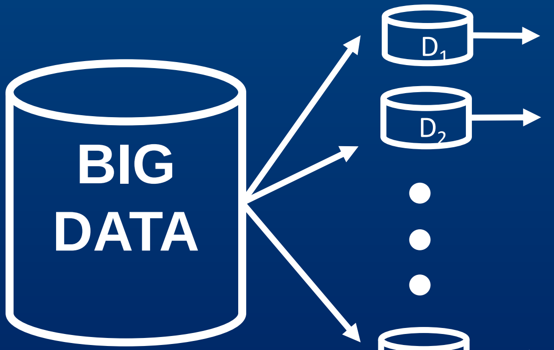
Hbase
Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System

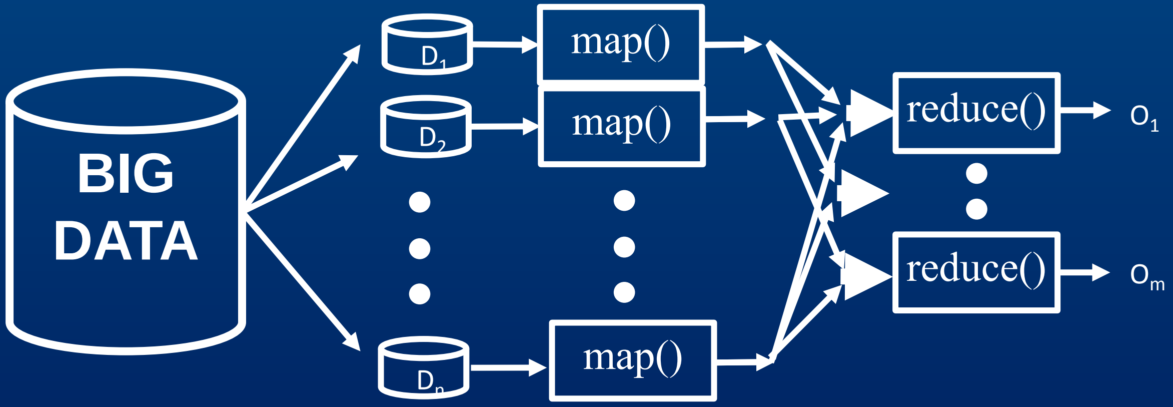
Hadoop Distributed File System (HDFS):

- partition data across cluster
- no schema so easy to add data



Map/Reduce Framework:

- Distribute computation to data
- Map – Hadoop Shuffle – Reduce



Map Shuffle Reduce

Hadoop (By Design):

- No schema, no index Pros :

  scalable, fault tolerant, fast loading,

- Cons:

  requires programming

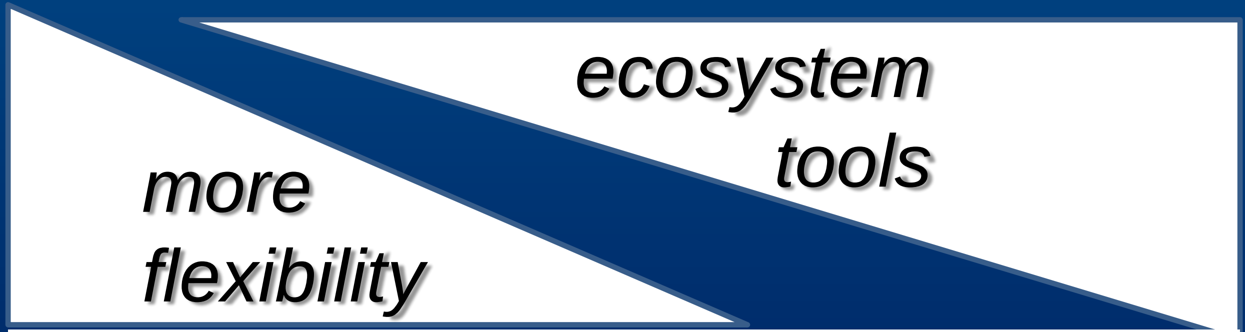
  no help with hierarchical relations

  no transaction reliability

DBMS:	Hadoop:
Structured data	Unstructured & growing
Transactional	Pass through data
SQL	Flexible mapping

The original functional view:

DBMS:	**Hadoop:**
Structured data	Unstructured & growing
Transactional	Pass through data
SQL	Flexible mapping



more flexibility

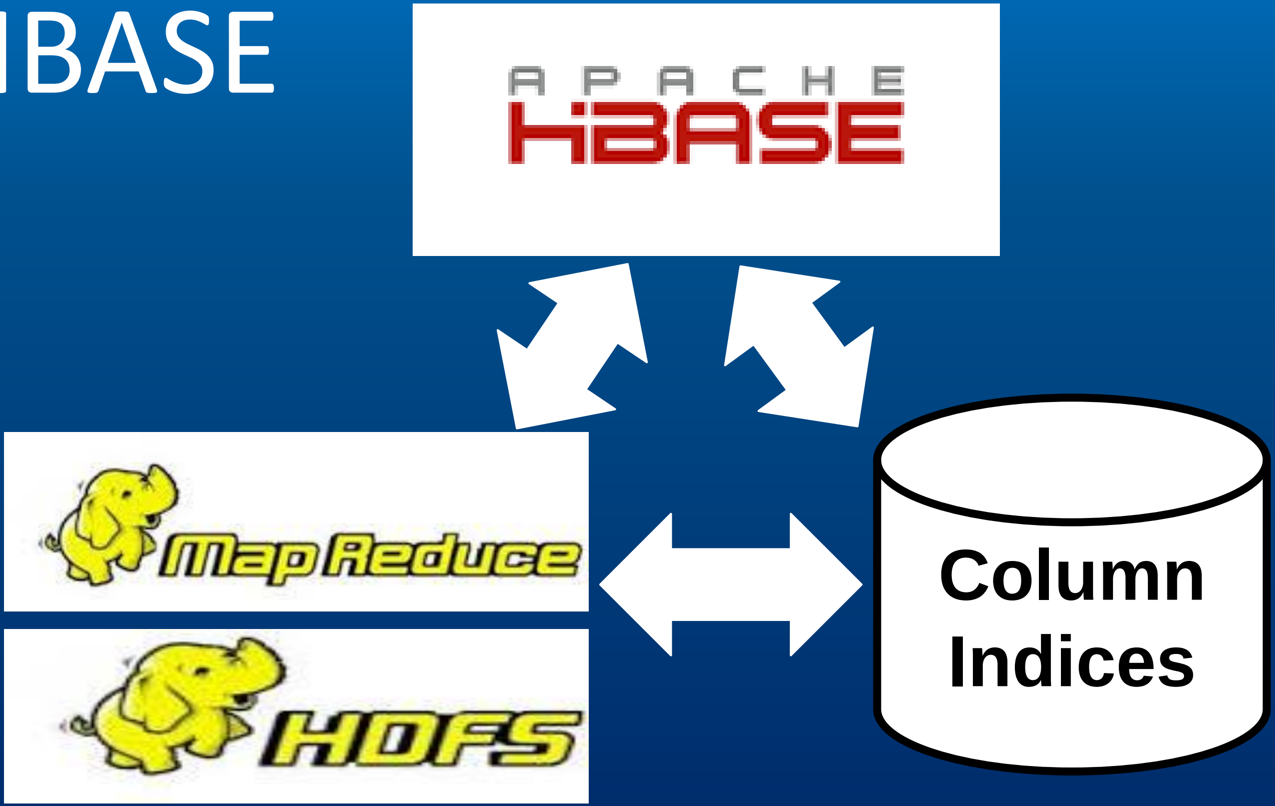
ecosystem tools

The functional trend – overlap & coexistence

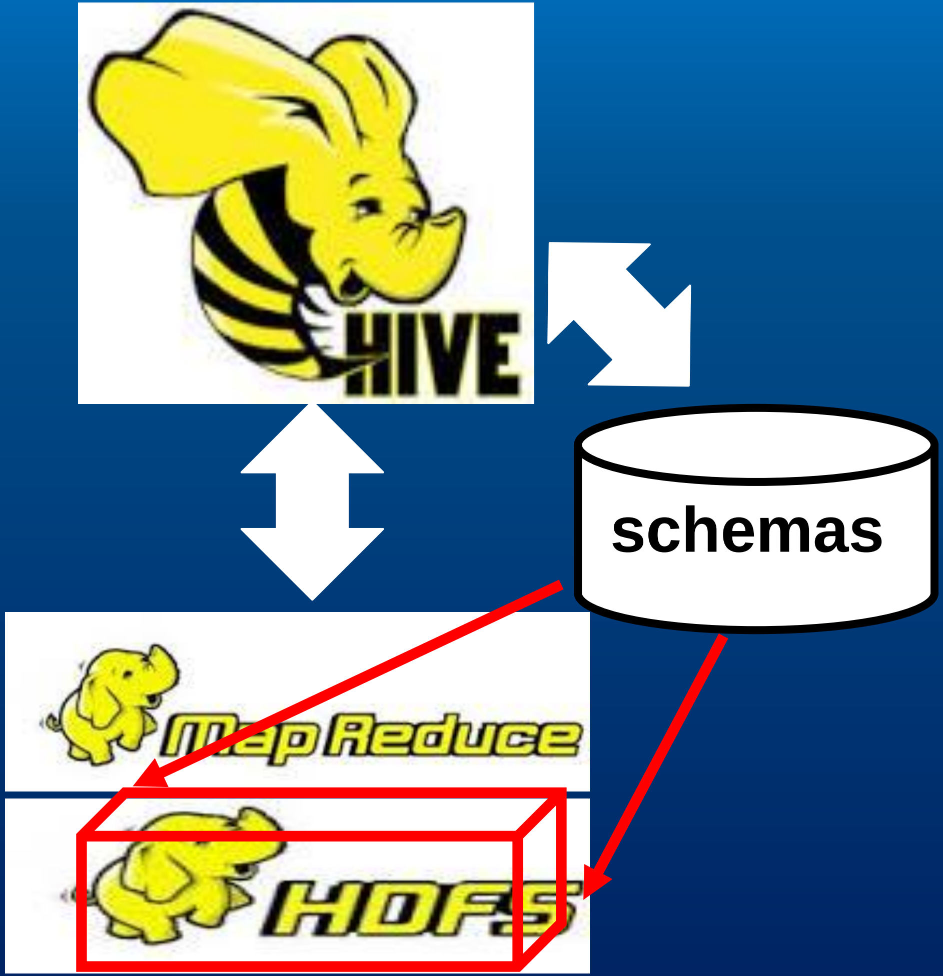
Ecosystem Tools

- query processing
- schema information
- utilities

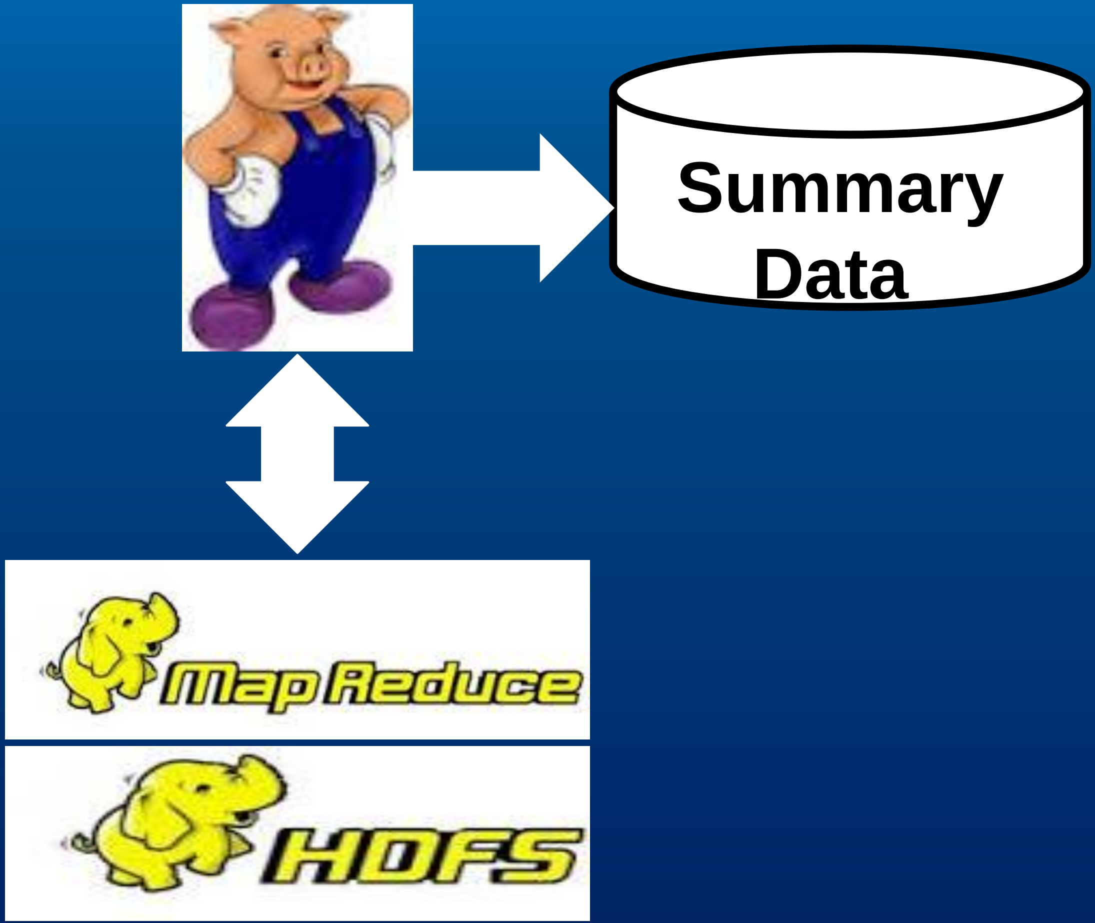
HBASE



HIVE

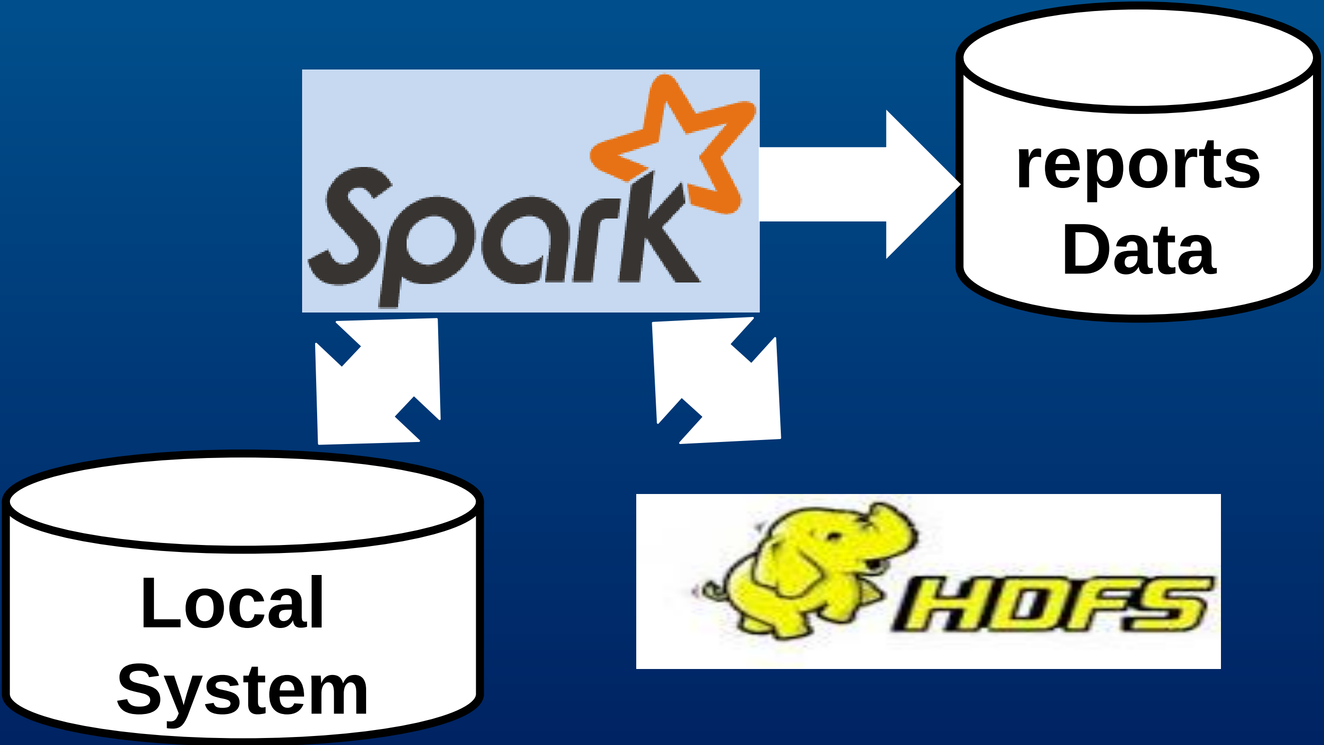


PIG



Summary Data

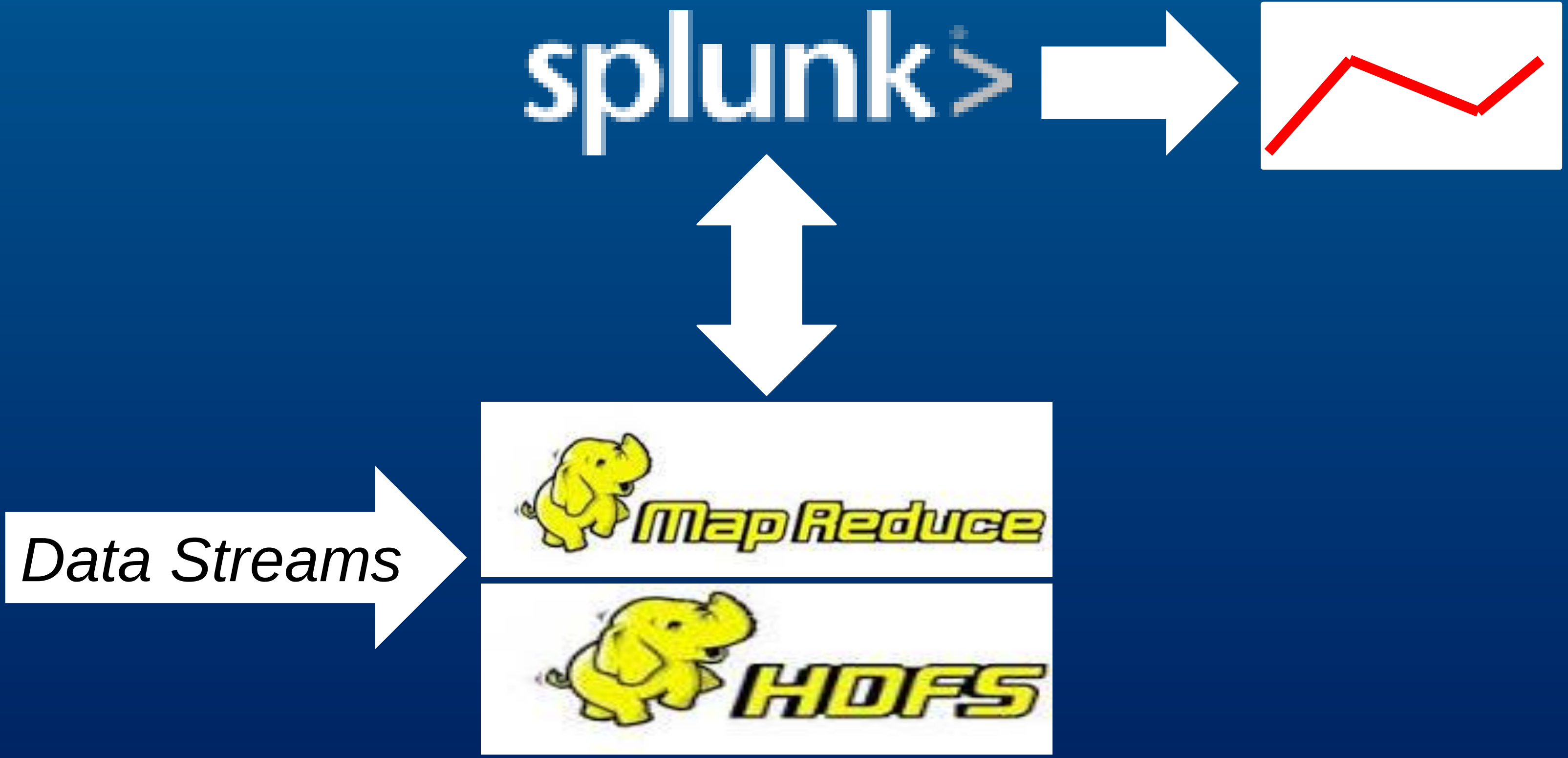
SPARK



reports Data

Local System

SPLUNK



Data Streams

- end

Big Data Analytics

What is Analysis?

Kinds of Analysis

Kinds of Analysis

- Query Processing

Kinds of Analysis

- Query Processing
- Summary Statistics

Kinds of Analysis

- Query Processing
- Summary Statistics
- Exploration

Kinds of Analysis

- Query Processing
- Summary Statistics
- Exploration
- Modeling

Query Processing

- SQL:

  select rows,

  project columns,

  join, group, sort, etc..

- Large queries need optimization

Descriptive Statistics

- Data Characteristics:

  sum, mean, variance,

  max, min, percentiles, etc…

- Often by groups

Exploratory Analysis

- Interactive
- Iterative
- Using samples

Sampling for Analysis

Sampling for Analysis

- Descriptive Statistics sample well

  sum, mean, max

Sampling for Analysis

- Descriptive Statistics sample well

  sum, mean, max

- Some queries do not sample well:

  joins, distinct, not in

Sampling for Analysis

- Descriptive Statistics sample well

  sum, mean, max, …
- Some queries do not sample well:

  joins, distinct, not in
- But some distribute well

  distinct, not in

Big vs Scaling

- Big Data Analysis is more than analysis with scaling:

  different tools,

  different questions & processing

- end