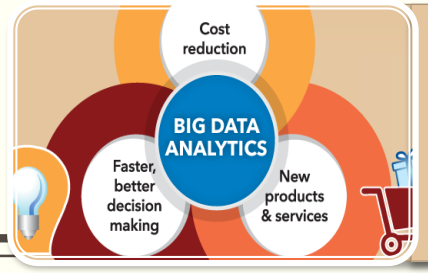
ภาษาการเขียนโปรแกรม 02-212-213

Week9

อ.ธิดาวรรร คล้ายศรี

บทที่3



Big Data Analytics

- Data Analytics Life Cycle
- Big data concepts & Hadoop ecosystem

3.1 Data Analytics Lifecycle (วงจรชีวิตการวิเคราะห์ข้อมูล)

3.1.1 Describing Problem (กำหนดปัญหา)

3.1.2 Identifying Data Sources (ระบุแหล่งข้อมูล)

3.1.3 Preparing Data (การเตรียมความพร้อมข้อมูล)

3.1.4 Planing Model (วางแผนโมเดล)

3.1.5 Building Model (สร้างโมเดล)

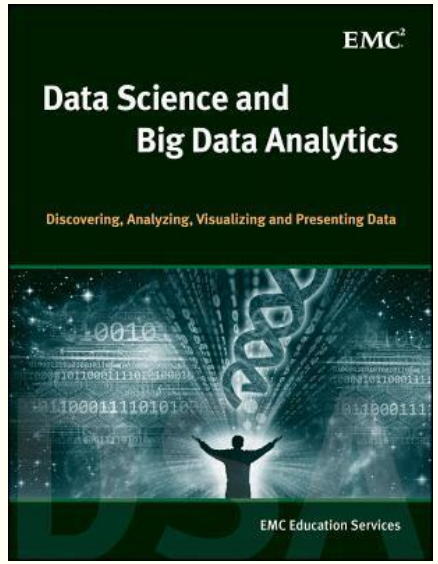
3.1.6 Visaulising Results (นำเสนอข้อมูลวิเคราะห์)

3.2 Big Data Concepts (แนวความคิดในเรื่อง big data)

3.3 Introductory Hadoop and Its Ecosystem (บทนำและวงชีวิตของฮาดูพ)

Source of the content:

EMC slides from the "**Data Science and Big Data Analytic**" text book by EMC Education Services



Part 3: 3.3 Introductory Hadoop & Its Ecosystem

บทนำฮาดูปและวงจรชีวิตของฮาดูป

A Technology for Big Data Analytics: Using Hadoop

What do we Mean by Hadoop

- A framework for handling big data
  - An implementation of the MapReduce paradigm
  - Hadoop glues the storage and analytics together and provides reliability, scalability, and management

Two Main Components

Storage (Big Data)

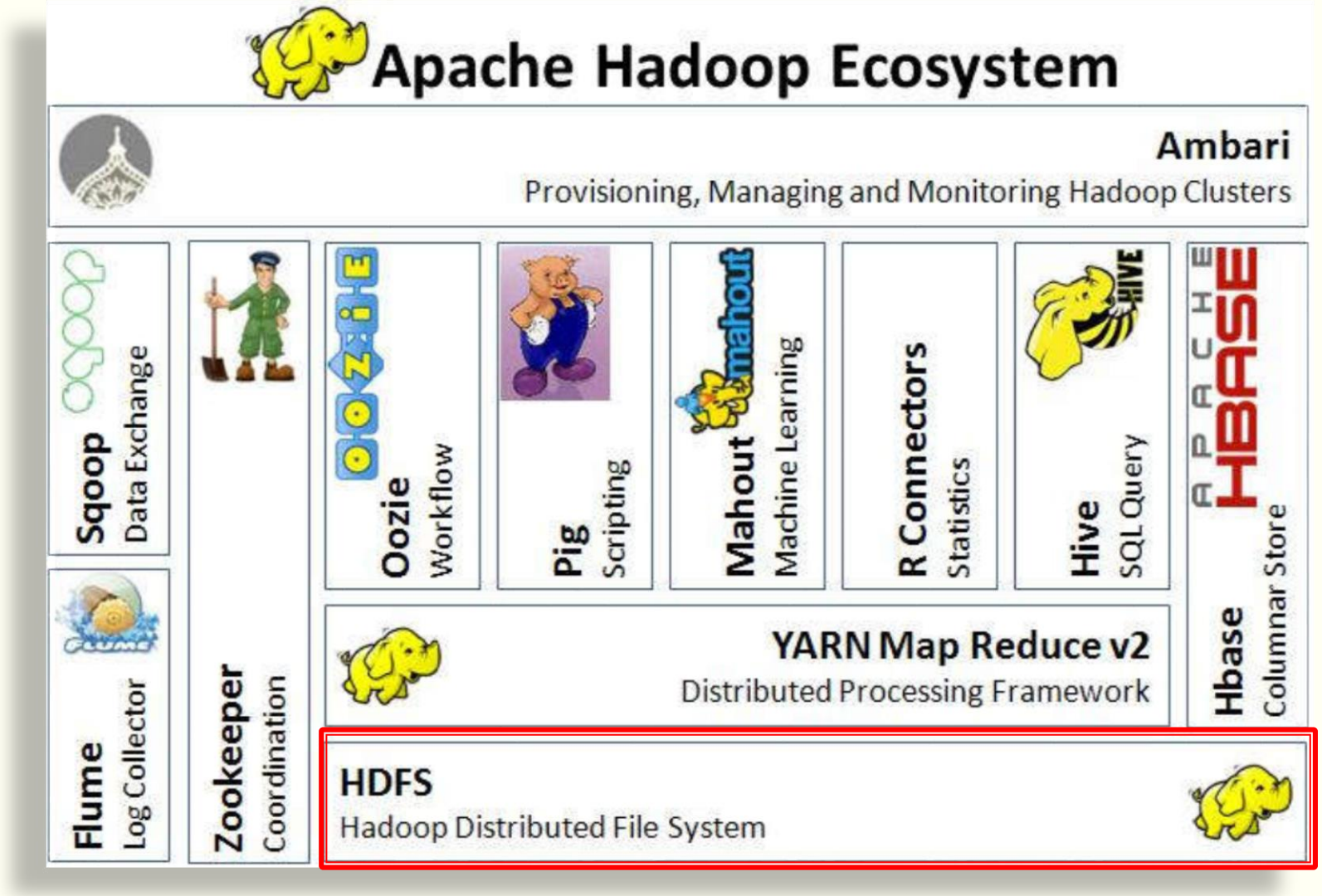
- HDFS – Hadoop Distributed File System
- Reliable, redundant, distributed file system optimized for large files

MapReduce (Analytics)

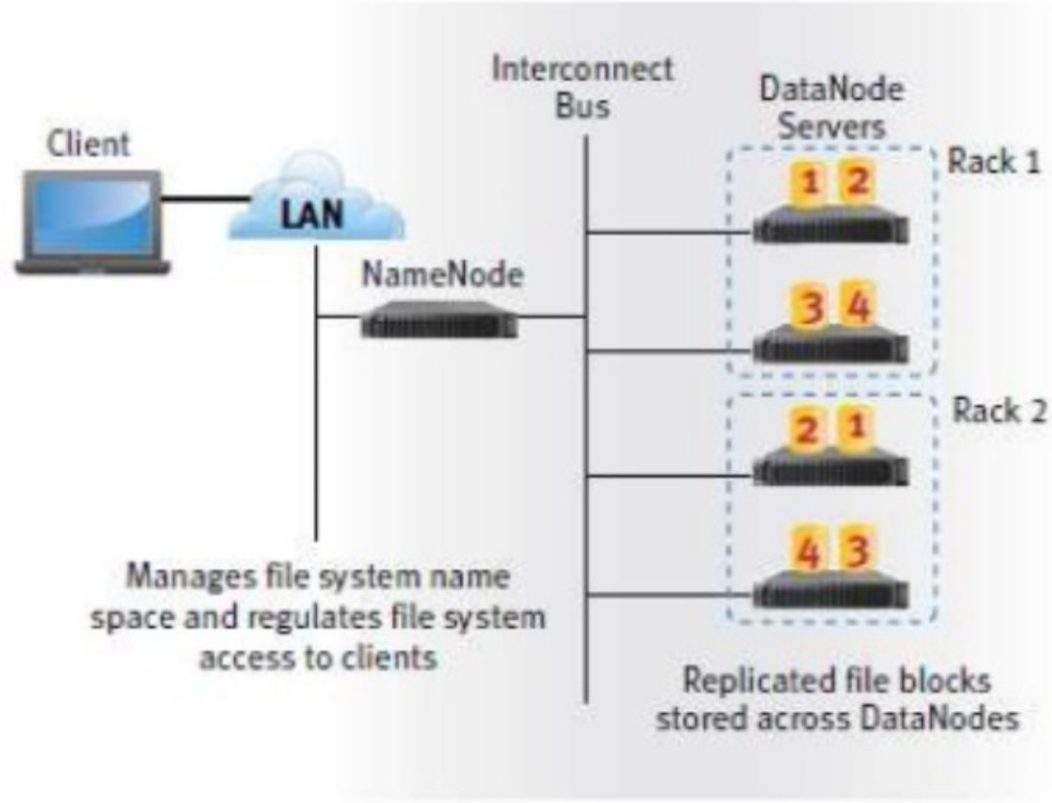
- Programming model for processing sets of data
- Mapping inputs to outputs and reducing the output of multiple Mappers to one (or a few) answer(s)

EMC² PROVEN PROFESSIONAL

Hadoop Ecosystem



Hadoop Distributed File System

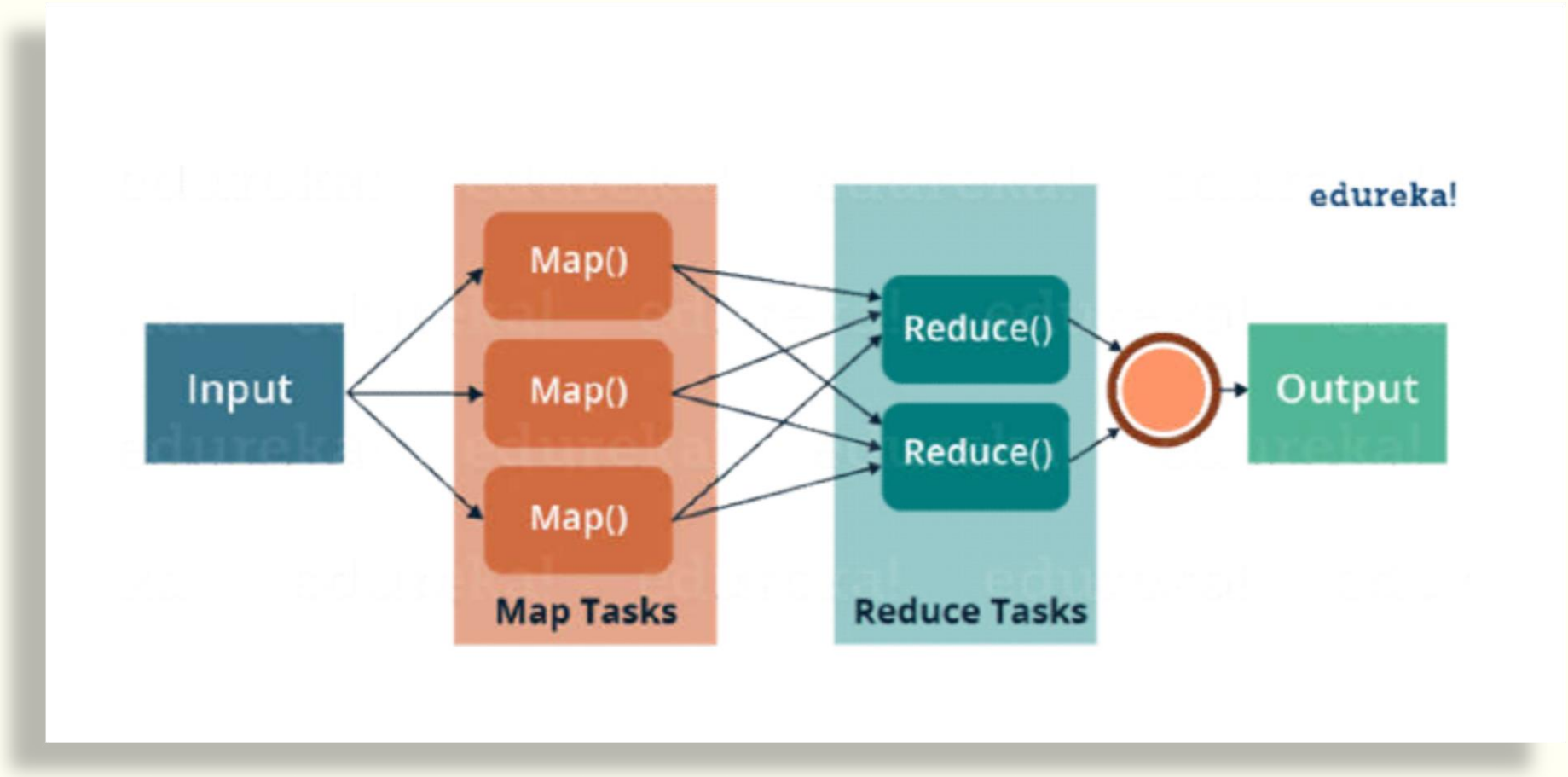


MapReduce คืออะไร

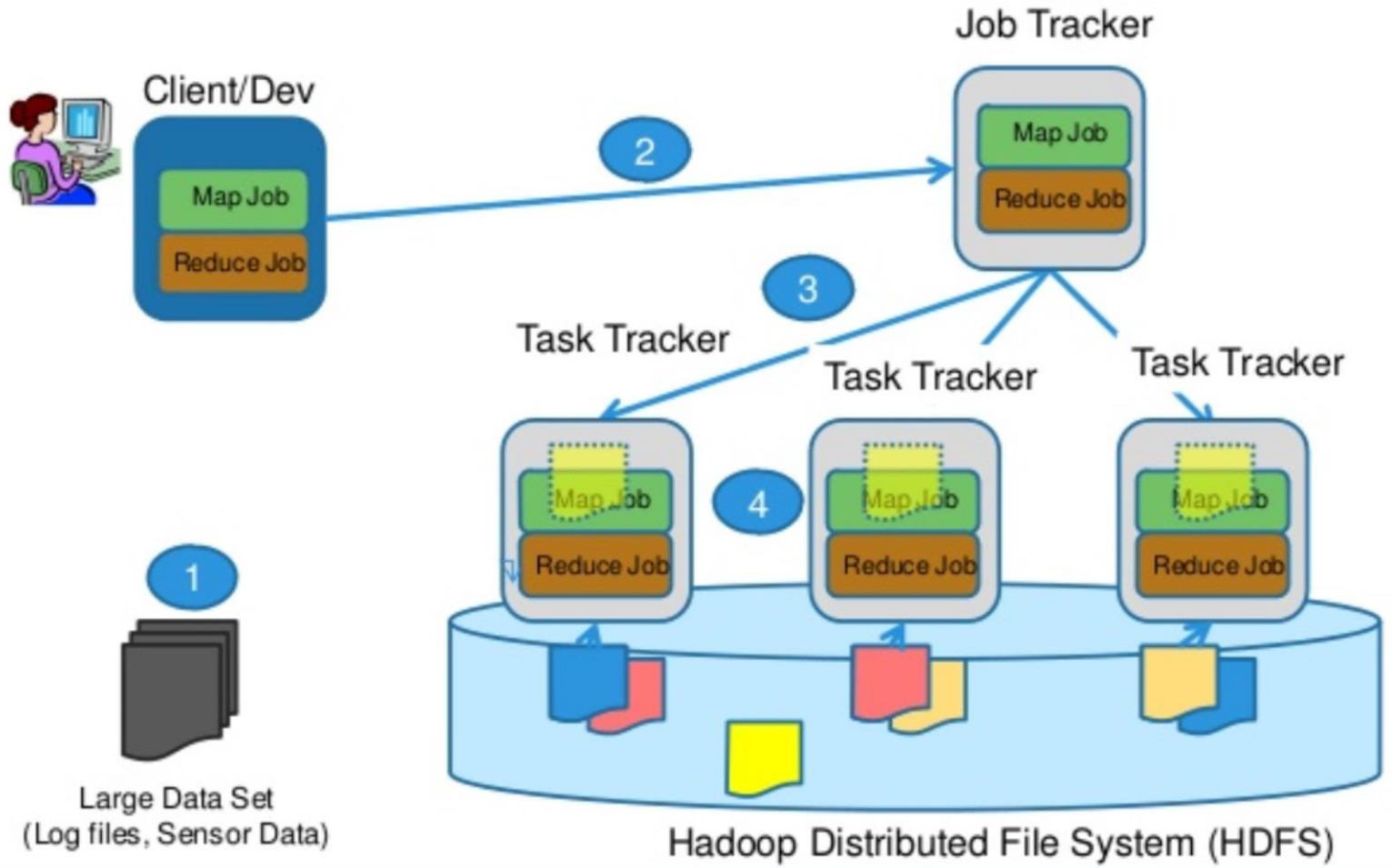
Google released a paper on MapReduce technology in December, 2004. This became the genesis of the Hadoop Processing Model. So, MapReduce is a programming model that allows us to perform parallel and distributed processing on huge data sets.

Google

หลักการ ของ MapReduce

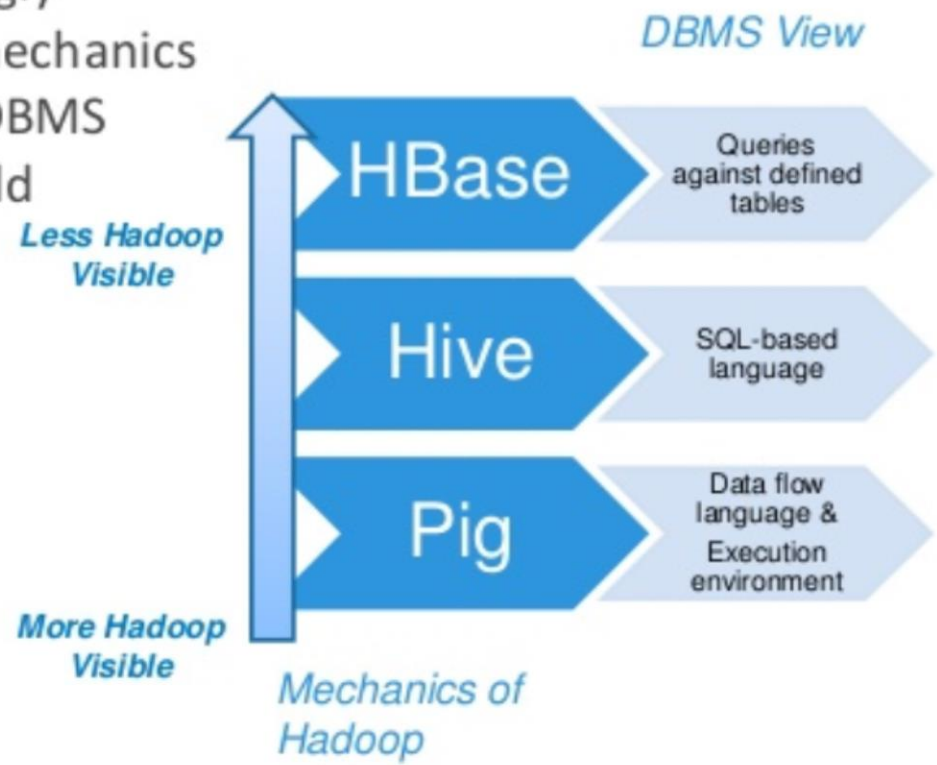


MapReduce and HDFS

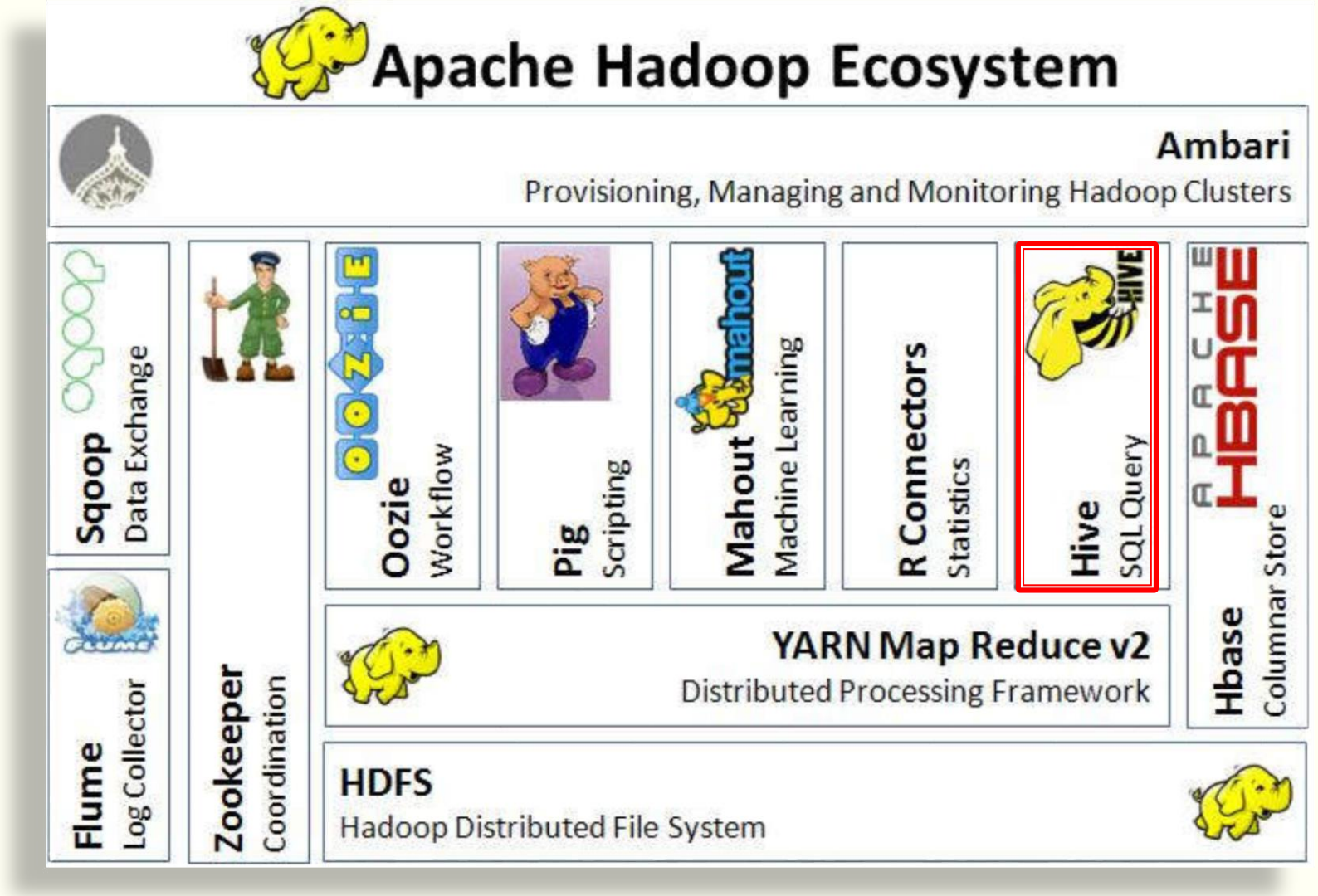


Components of Hadoop

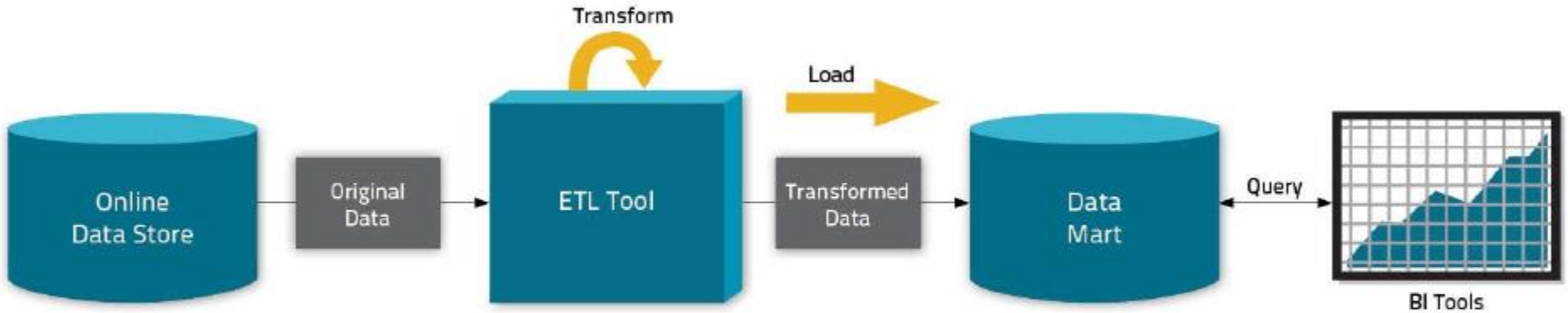
- As you move from Pig to Hive to HBase, you are increasingly moving away from the mechanics of Hadoop and get an RDBMS view of the Big Data world



Hadoop Ecosystem



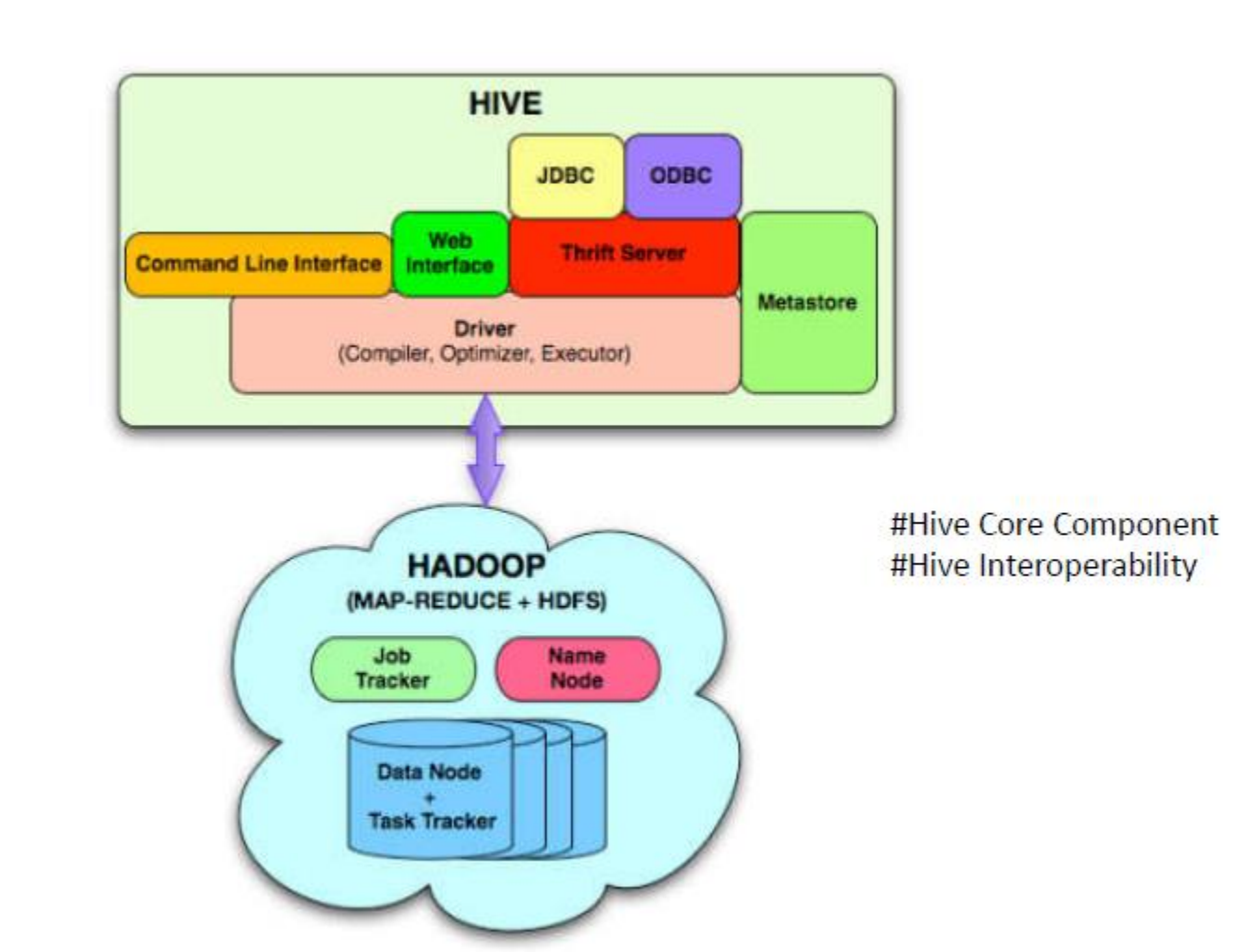
Hadoop data Challenge with data processing :



ตัวอย่าง Applications สำหรับ Hive

- Summarization
  - Eg: Daily/Weekly aggregations of impression/click counts
  - Complex measures of user engagement
- Ad hoc Analysis
  - Eg: how many group admins broken down by state/country
- Data Mining (Assembling training data)
  - Eg: User Engagement as a function of user attributes
- Ad Optimization

Hive’s components



#Hive Core Component
#Hive Interoperability