cloudera®

CAP - Developing with Spark and Hadoop

201509

cloudera®

Introduction

Chapter 1

Course Chapters

1	**Introduction**	**Course Introduction**
2	Introduction to Hadoop and the Hadoop Ecosystem	Introduction to Hadoop
3	Hadoop Architecture and HDFS	
4	Importing Relational Data with Apache Sqoop	Importing and Modeling Structured Data
5	Introduction to Impala and Hive	
6	Modeling and Managing Data with Impala and Hive	
7	Data Formats	
8	Data Partitioning	
9	Capturing Data with Apache Flume	Ingesting Streaming Data
10	Spark Basics	Distributed Data Processing with Spark
11	Working with RDDs in Spark	
12	Aggregating Data with Pair RDDs	
13	Writing and Deploying Spark Applications	
14	Parallel Processing in Spark	
15	Spark RDD Persistence	
16	Common Patterns in Spark Data Processing	
17	Spark SQL and DataFrames	
18	Conclusion	Course Conclusion

Chapter Topics

Introduction | **Course Introduction**

- **About This Course**
- About Cloudera

Course Objectives

During this course, you will learn

- **How the Hadoop Ecosystem fits in with the data processing lifecycle**
- **How data is distributed, stored and processed in a Hadoop cluster**
- **How to use Sqoop and Flume to ingest data**
- **How to process distributed data with Spark**
- **Best practices for data storage**
- **How to model structured data as tables in Impala and Hive**
- **How to choose a data storage format for your data usage patterns**

Chapter Topics

Introduction | **Course Introduction**

- About This Course
- **About Cloudera**

About Cloudera (1)

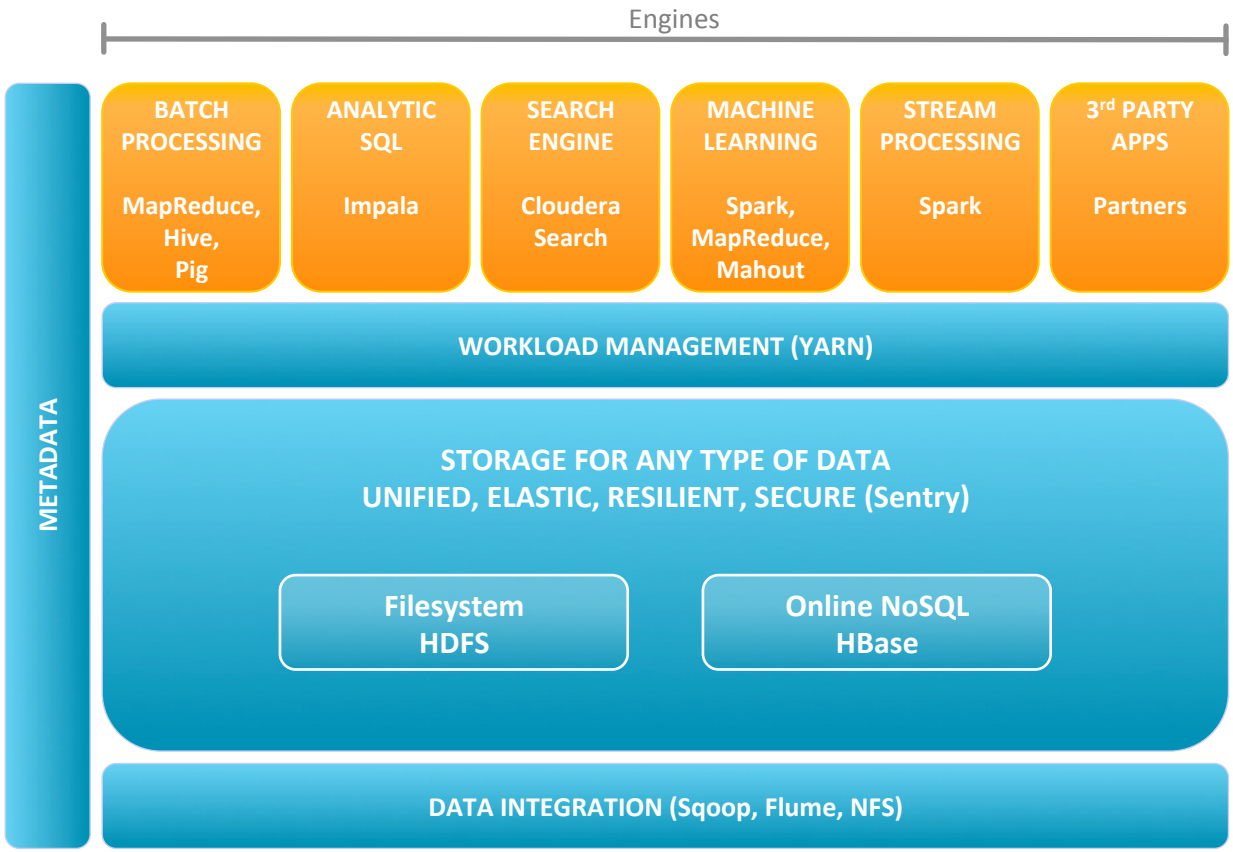
- **The leader in Apache Hadoop-based software and services**
- **Founded by leading experts on Hadoop from Facebook, Yahoo, Google, and Oracle**
- **Provides support, consulting, training, and certification for Hadoop users**
- **Staff includes committers to virtually all Hadoop projects**
- **Many authors of texts on Apache Hadoop projects, including**
  - Lars George, Amandeep Khurana, Uri Laserson, Sean Owen, Sandy Ryza, Ben Spivey, Kathleen Ting, Tom White, and Josh Wills

About Cloudera (2)

- **Customers include many key users of Hadoop**
  - Allstate
  - AOL Advertising
  - Box
  - CBS Interactive
  - eBay, Experian
  - Groupon
  - National Cancer Institute
  - Orbitz
  - Social Security Administration
  - Trend Micro
  - Trulia
  - US Army

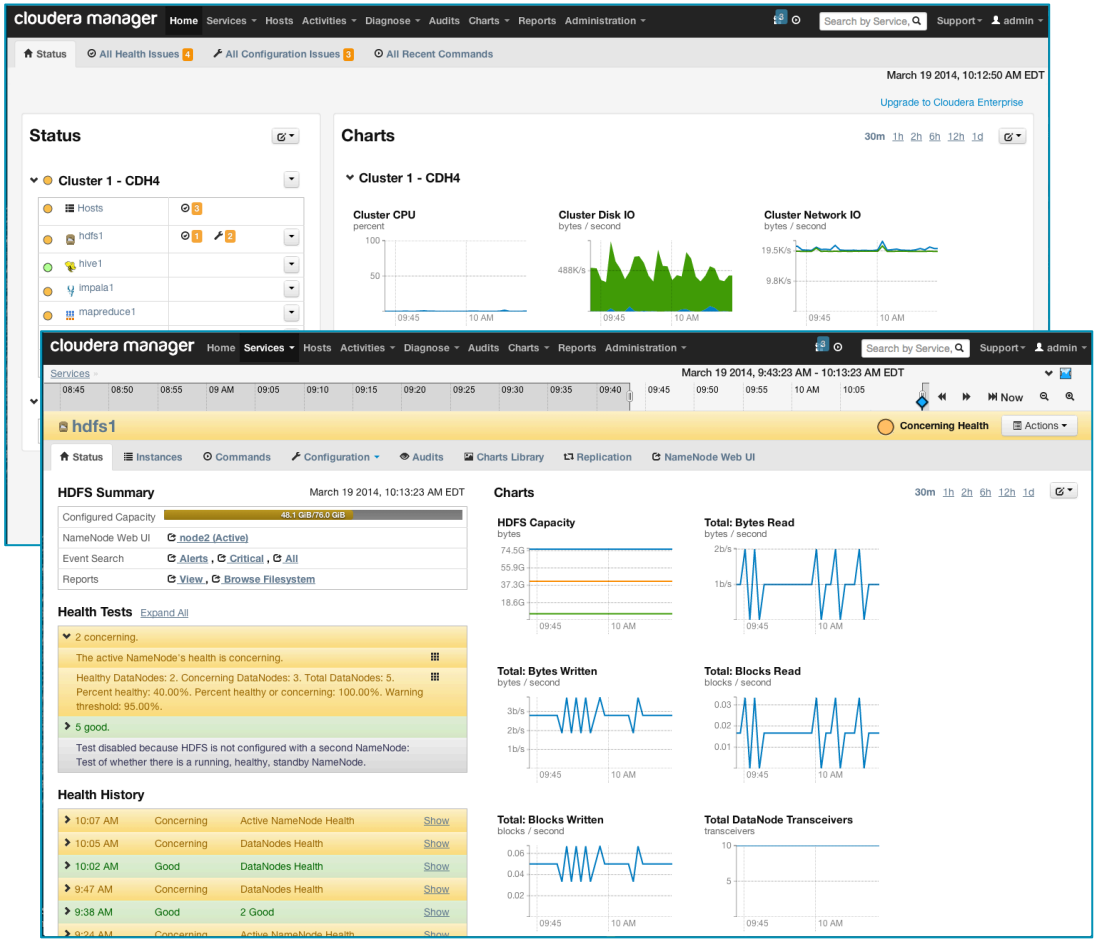
CDH (Cloudera's Distribution including Apache Hadoop)

- 100% open source, enterprise-ready distribution of Hadoop and related projects
- The most complete, tested, and widely-deployed distribution of Hadoop
- Integrates all the key Hadoop ecosystem projects
- Available as RPMs and Ubuntu, Debian, or SuSE packages, or as a tarball



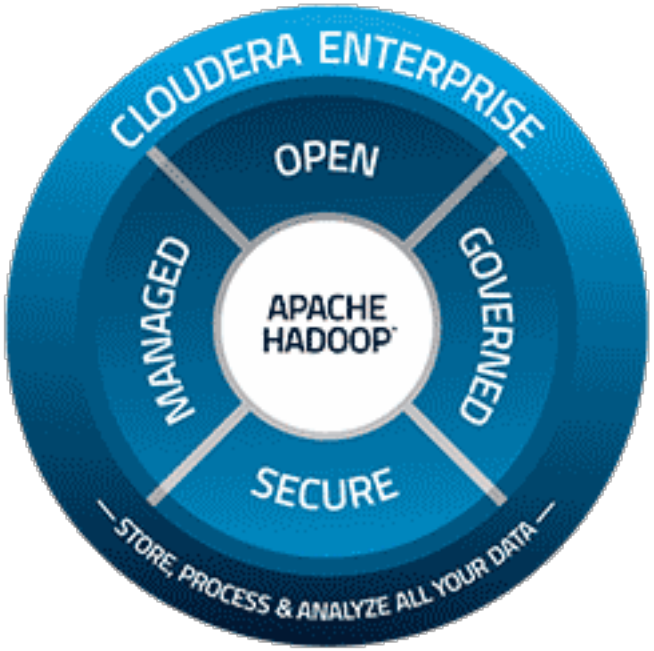
Cloudera Express

- **Cloudera Express**
  - Completely free to download and use
- **The best way to get started with Hadoop**
- **Includes CDH**
- **Includes Cloudera Manager**
  - End-to-end administration for Hadoop
  - Deploy, manage, and monitor your cluster



Cloudera Enterprise

- **Cloudera Enterprise**
  - Subscription product including CDH and Cloudera Manager
- **Includes support**
- **Includes extra Cloudera Manager features**
  - Configuration history and rollbacks
  - Rolling updates
  - LDAP integration
  - SNMP support
  - Automated disaster recovery



- **Extended capabilities with Cloudera Navigator subscription**
  - Event auditing, metadata tagging capabilities, lineage exploration