INTRODUCTION

Introduction

- **The Apache Software Foundation**
- Apache Hadoop and Big Data Revolution
- Characteristics of Big Data
- Apache Hadoop Ecosystem

ASF

- An organization for a number of Open Source projects
- Since 1999, ~150 projects, ~2000 committers
- 50% of all open source downloads are Apache projects
- *The Apache Way*: Meritocracy in action

ASF

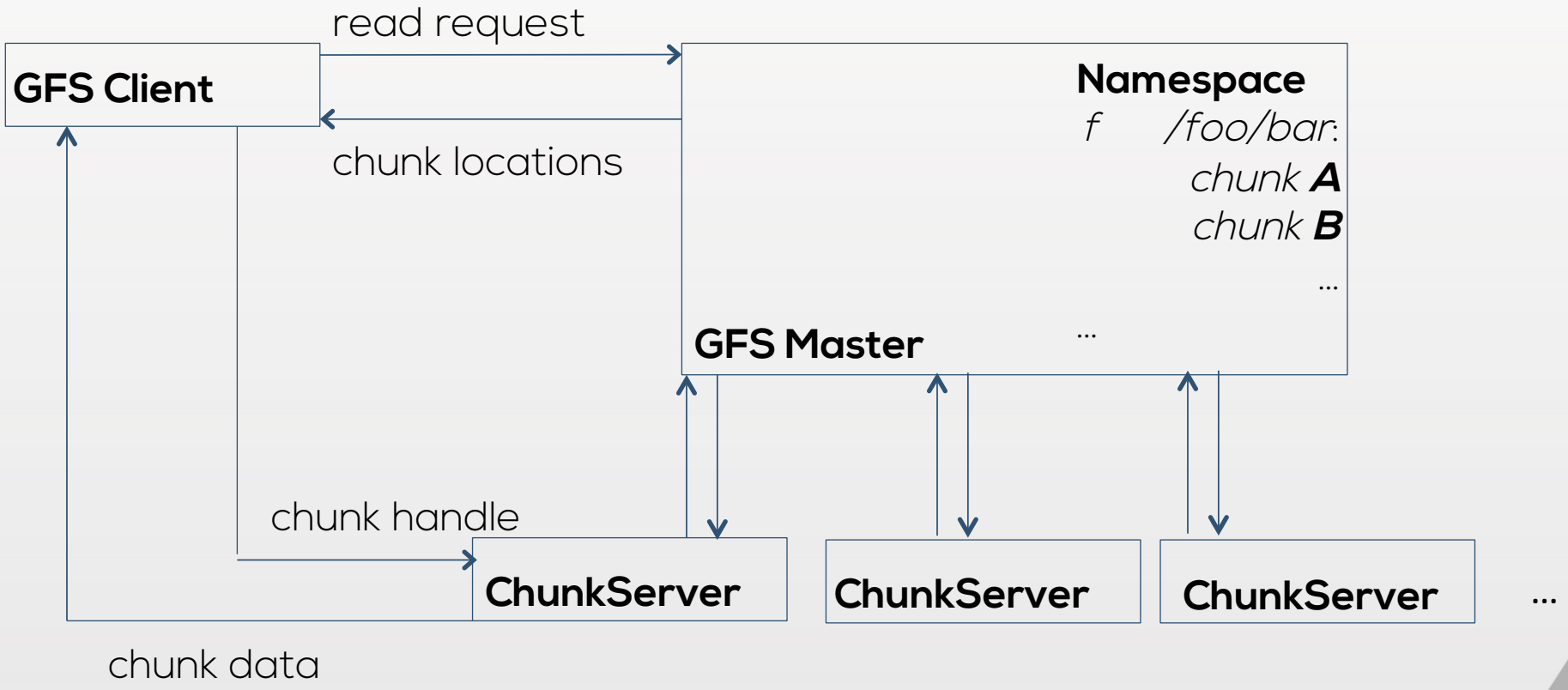
- All ASF projects, and more, are licensed with “Apache License, version 2.0”, and it allows users:
  - To use the software freely
  - To redistribute and sublicense it with or without changing its code
- Big Data Ecosystem is almost 100% Open Source, and the relevant components are usually independent Apache Projects

Introduction

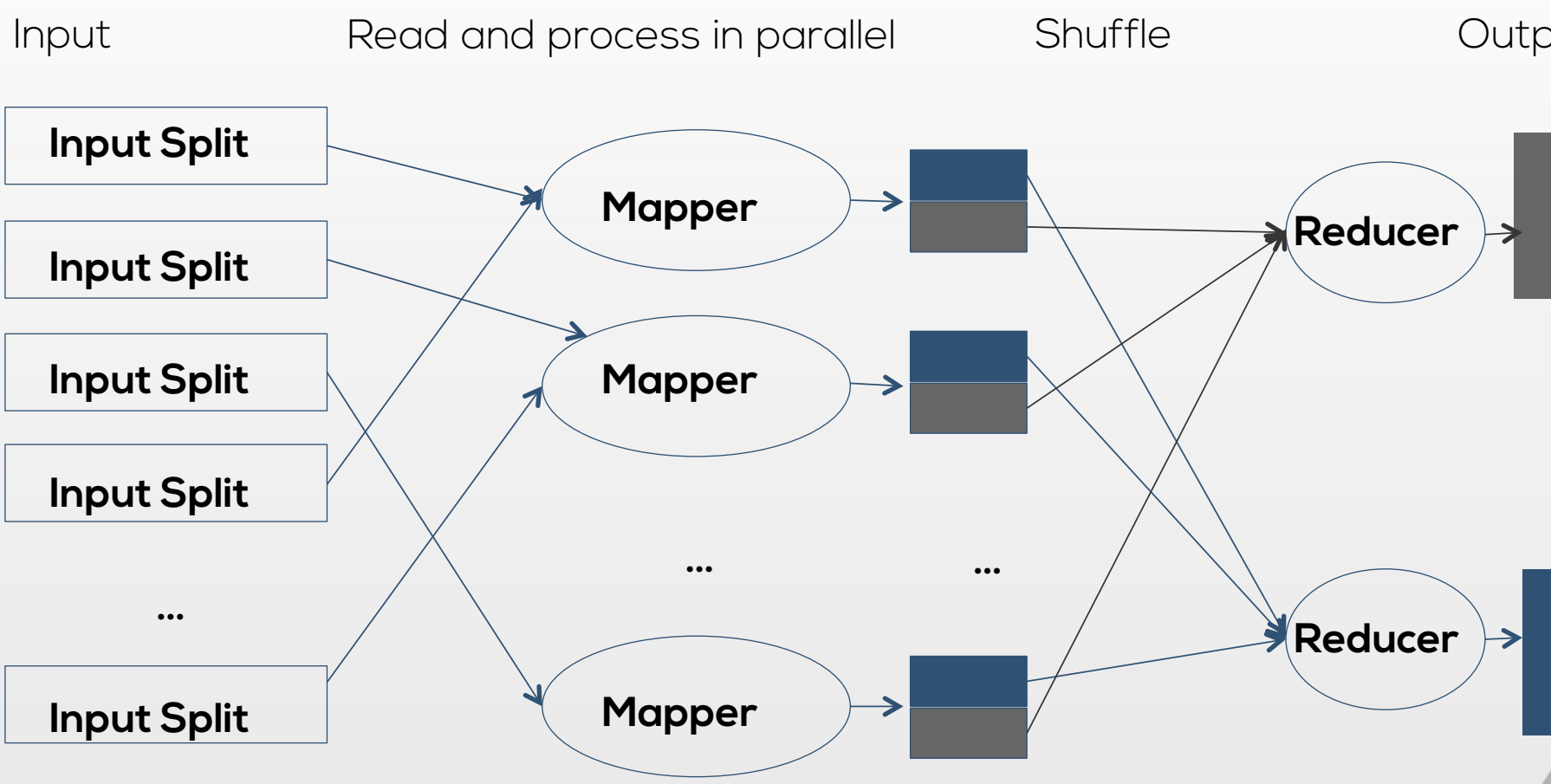
- The Apache Software Foundation
- **Apache Hadoop and Big Data Revolution**
- Characteristics of Big Data
- Apache Hadoop Ecosystem

Apache Hadoop

- Implemented based on **Google's GFS** and **MapReduce** papers, **Apache Hadoop** was initially a component of another Apache project, an open source search engine, **Nutch**
  - It served to the purpose of crawling, storing, and indexing web pages in a distributed fashion, initially
- Distributed computing part of Nutch had received an instant interest and then became a project on its own, and named **Hadoop**



GFS Architecture



Google's MapReduce

Apache Hadoop

- On a **cluster of servers**, the two core components of this software, the **Hadoop Distributed File System** and the **MapReduce** framework allowed
  - Storing data in a **distributed** and **fault tolerant** fashion (HDFS)
  - **Parallel** and **distributed** processing of data stored in this distributed file system (MapReduce)

Apache Hadoop

- In a very short time Apache Hadoop extended beyond its original purpose, and used by many organizations, due to:
  - HDFS's **scalability**, **fault tolerance**, being able to run on **commodity hardware**
  - MapReduce's simple programming model, ability to **move computation to where data partitions reside**, and its **inherent characteristic of being general-purpose** (both in terms of **flexibility in input**, and the **applicability to a broad class of data processing problems**)

Big Data Revolution

- It is the Apache Hadoop software that created the Big Data Buzz, mainly because
  - it is an **open source software**, and its direction is decided and foreseen by the community
  - it provides an **affordable** and **scalable** way of storing **massive amounts of data**
  - it provides a simple, yet generic-enough way of processing such data:
    - Generic in the data processing applications it supports
    - Genetic in the type of data it can process

Introduction

- The Apache Software Foundation
- Apache Hadoop and Big Data Revolution
- **Characteristics of Big Data**
- Apache Hadoop Ecosystem

Characteristics of Big Data

- Hadoop's being
  - **scalable**, **affordable**, **fault-tolerant** in storage,
  - **flexible** in data types
  - **generic** in processing of data

  allowed us to **collect**, **store** and **analyze data** in such a way that if it wouldn't be possible otherwise (feasible, if you want)

Characteristics of Big Data

- We call such data Big Data; characterized by its:
  - **Massive** amount
  - **Growing** in amount
  - Being in **various formats**, collected from **various channels**

Characteristics of Big Data

- Examples of such data include
  - Messages and emails that people send to each other
  - Logs and reviews people provide on a web site
  - Collection of Twitter status updates
  - Call data
  - Web Server Log
  - Location-tracking data
  - Medical records
  - …

Some Technical Traits

- Some traits of Big Data:
  - Data is **split into multiple partitions** residing in distant disks
  - The dataset is **a huge collection of record**s
  - **Moving the data around** is **intolerably expensive**
  - **Communication effectiveness** (reducing the amount of data moving between nodes) will be key to **algorithms for Big Data** Processing

Introduction

- Apache Hadoop and Big Data Revolution
- Characteristics of Big Data
- **Apache Hadoop Ecosystem**

Apache Hadoop Ecosystem

- At the core of processing Big Data, we have Hadoop (with the distributed storage HDFS, and cluster resource manager YARN)
- Around this core there is a large ecosystem of tools, which might serve several purposes:
  - **Distributed Programming Engines** that implement the MapReduce-like processing model (MapReduce, Spark, …)
  - **Supplementary Tools** to complete the Big Data procesing pipeline (Flume, Sqoop, …)
  - **High Level Abstractions** that make working with Hadoop easier, standardized, and integrated to popular tools (Pig, Hive, …)
  - **Libraries** as collections of data processing algorithms (graph processing libraries, machine learning libraries such as Mahout, …)

Apache Hadoop Ecosystem

- Examples of ecosystem tools:
  - NoSQL databases like **Apache HBase** and **Accumulo**
  - Data collection tools like **Apache Flume** and **Sqoop**
  - **Apache Hive:** Data warehousing on Hadoop with an SQL-compliant query language
  - **Apache Pig**: A higher level abstraction to MapReduce with an easy to use dataflow language called Pig Latin, supporting various MapReduce usage patterns and relational operations
  - **Apache Mahout**: A library of scalable Machine Learning algorithms
  - **Apache Spark**: An alternative to MapReduce and Stream Processing, surrounded by machine learning, graph procesing, SQL querying components

Apache Hadoop Ecosystem

- This is a very large ecosytem with many related but independent projects with their own development processes and release cycles
- It is nontrivial
  - to make use of all such tools (**learning**),
  - to understand how they should be combined in order to achieve a goal (**solution architecting**),
  - to keep these independent tools up to date, integrated, and operational (**administrating**)

Apache Hadoop Ecosystem

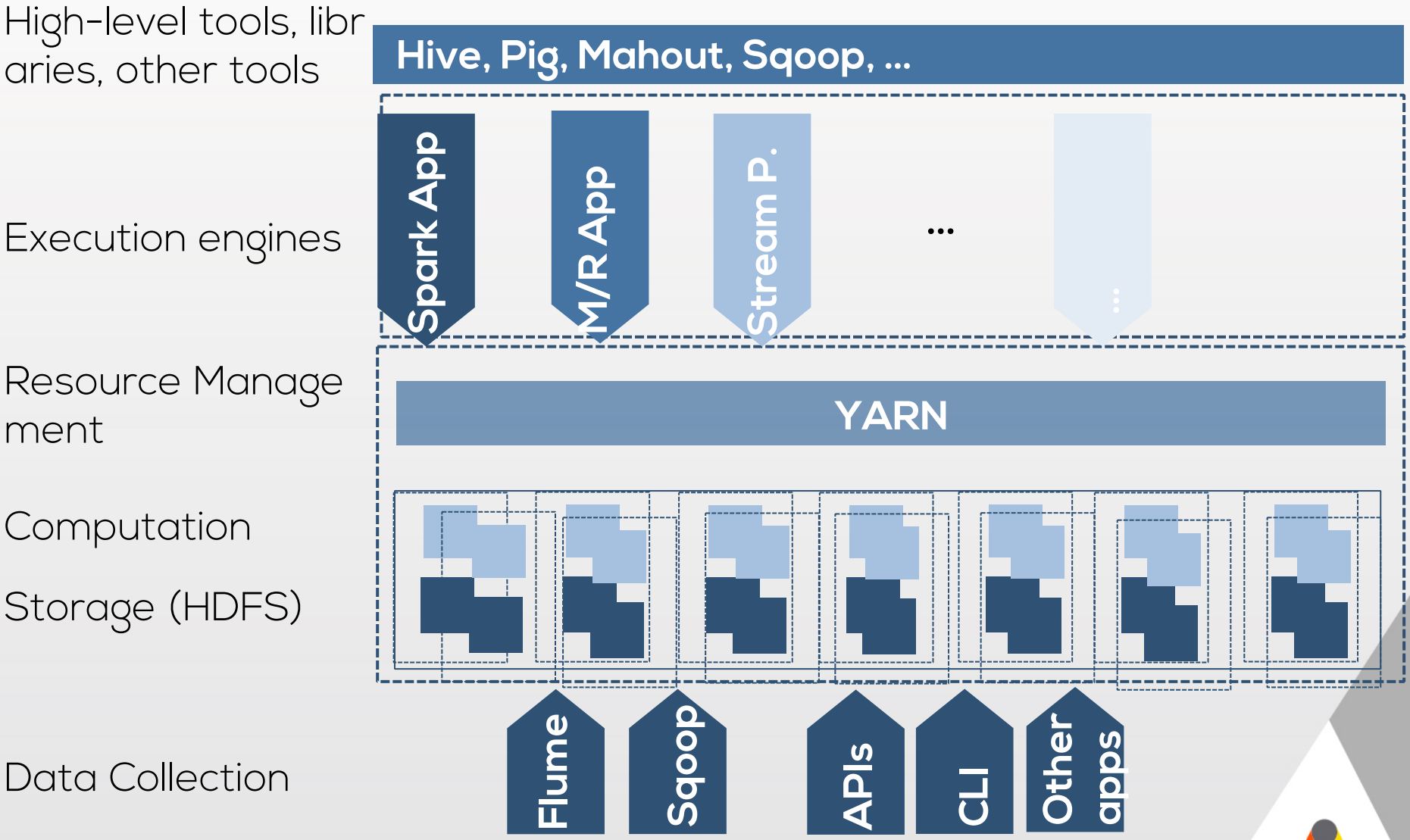
- There are other Apache projects, such as Bigtop, with a sole purpose of testing and integrating Hadoop and friends; as well as companies offering a Big Data processing stack
- Such collections of software are usually referred to as an **Apache Hadoop distribution**, or **Big Data distribution**
- There are also commercial Big Data distribution offerings, such as **CDH of Cloudera** and **HDP of Hortonworks**, additionally easing the users jobs to utilize and administrate a Big Data clusters, and develop solutions on top of it
- **AnalyticsCenter Faculty** deliberately stays vendor-independent, and the training VMs you use are Bigtop clusters.

Apache Hadoop Ecosystem

- A typical Hadoop cluster would include
  - HDFS as the common distributed storage
  - HBase/Accumulo as the low-latency NoSQL storage
  - YARN as the cluster management software (scheduling and allocating cluster resources to different applications), with compute nodes colocated with the nodes making the HDFS
  - MapReduce and Spark applications running in a distributed fashion
  - Server components of higher level tools, such as HiveServer for submitting Hive Queries, Oozie server for workflow submission, etc.

Apache Hadoop Ecosystem

- Client applications typically interact with the cluster via
  - Libraries like Apache Mahout
  - Pig Scripts via Apache Pig
  - SQL queries via Apache Hive
  - MapReduce (Java) or Spark Applications
  - Flume agents
  - Sqoop commands
  - …



A typical Hadoop cluster

Summary

- The Apache Software Foundation
  - ASF organization and its role in the Big Data world
- Apache Hadoop and Big Data Revolution
  - The core technology powering Big Data, and its evolution
- Characteristics of Big Data
  - When we refer to data as "Big"
- Apache Hadoop Ecosystem
  - The technologies around Apache Hadoop to make processing Big Data possible

Introduction

End of Chapter