Data Centric Systems and Networking Wrapup

Eiko Yoneki

University of Cambridge Computer Laboratory
Big Data Analytics Stack

- Streaming Processing
- Query/Scripting Language
- Machine Learning
- Graph Processing
- Execution Engine
- Resource Manager
- Database
- Storage
Hadoop Big Data Analytics Stack

[Diagram showing the Hadoop Big Data Analytics Stack with components such as Pig/Hive, Mahout, Giraph, MapReduce, YARN, HBase, and HDFS]
Spark Big Data Analytics Stack

Spark

Mesos, YARN

HBase

HDFS
Big Data Analytics Stack

**Language**
- Pig, Hive, Shark, Meteor, SCOPE, DryadLINQ

**ML + DL(?)**
- Mahout, MLBase, SystemML, Presto

**Streaming**
- Storm, S4, SEEP, Dstream, Naiad

**Graph Processing**
- Pregel, GraphLab, Bagel, GraphX, Giraph

**Engine**
- MapReduce, Dryad, Spark, Nephele/PACT, Hayracks, Percolator

**Resource Manager**
- Mesos, YARN

**Database**
- BigTable, Hbase, Dynamo, Cassandra, MongoDB, Voldemort

**Storage**
- HDFS, GFS
Large-scale Data Processing Technologies

- Distributed infrastructure or MC/HPC
  - Cloud (e.g. Infrastructure as a service)
- Storage
  - Distributed storage (e.g. Amazon S3)
  - HDFS, GFS, Dynamo - HBase, BigTable
- Data model/indexing
  - High-performance schema-free database
- Programming Model
  - Distributed processing (e.g. MapReduce, Pregel (BSP), or more complex data-flow programming)
- Operations on big data
  - Analytics – Realtime Analytics

R212 Focus: Data driven approach is key!
Data Centric Networking

- Which session’s topic did you enjoy the most?

- What topic do you want to explore if you go on research work?

- What do your future vision of data centric systems and networking?
Data Centric Systems and Networking

- That’s all!

- Thanks for working hard in this course!