

Final Project Preparation

AWS vs. Hadoop

AWS

- For more than big data
- More recently updated
- Requires money/school credit
- Allows you to use existing programs for data analysis

Hadoop

- Only big data
- Older
- Free
- Utilize novel programs for data analysis

Vs of Big Data

- Volume: How much?
- Variety: What kind?
- Velocity: How fast?
- Veracity: How trustworthy?
- Value: Why?

Data Streaming

- Real-time data
- Includes data generated by machine or IoT
- How do you store / extract it?
 - Kafka
 - Flume
 - Flink
- How do you analyze it?
 - Spark Streaming
 - Storm
 - Flink

Pig vs. Hive

Pig

- Uses a SQL-like language (Pig Latin)
- Slower
- Data = relations
- More flexible

Hive

- Uses a version of SQL (HiveQL)
- Quicker
- Data = tables
- Easier to use

MapReduce

- The original big data system
- Map – find the data you need and transform it
- Reduce – aggregate it

Questions?