Big Data Analytics

Georgia Tech

What is Big Data?

Caracterized by

- Volume
  - No specific threshhold, but typically several gigabytes (10_9), terabytes (10_{12} or petapbytes (10_{15})
- Velocity – the data are generated quickly
  - Facebook generates 600 TB of new data per day. [1]
- Variety – from multiple, often heterogeneous sources
- Variability – incomplete data, inconsistency within and between data sources
- Veracity – how can you trust the data you ingest?

A good operative definition: a data set that may not fit on a single hard disk and/or requires parallel computation to process in a reasonable amount of time. (In practice many "big data" sets measure in the gigabytes, which might actually fit on a single modern disk.)

[1] Pamela Vagata and Kevin Wilfong, *Scaling the Facebook data warehouse to 300 PB*

Applications of Big Data

- Web search
- Ad serving
- Multimedia analytics (image, video)
- Collaborative filtering (e.g., "customers who viewed this also viewed")
- Customer churn (identify customers likely to switch to a competitor in order to target special offers aimed at retention)
- Health care analytics
- Any sort of analytics application where the scale requires "big data" technology for reasonable performance.

Big data processing is typically done in batch mode. A new paradigm, fast data, has recently emerged in which data are processed in real-time, often in combination with some batch-mode processing. We'll focus on batch mode big data processing here, which is also typically a component of fast data systems.

Georgia Tech

Managing Big Data

The characteristics of big data lead to two primary technical challenges:

- storage, and
- parallel processing.

We'll explore these challenges in the context of a ubiquitous industry-standard solution: the Hadoop scalable distributed computing platform.

Georgia Tech

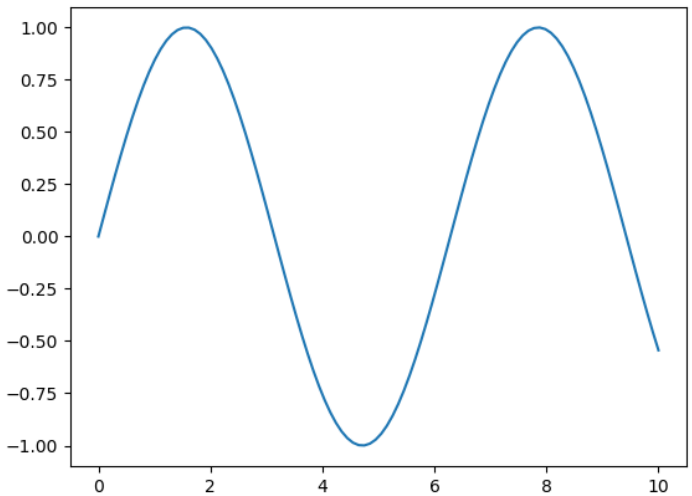
The Hadoop Platform

Hadoop is not a single software product, but an ecosystem of software tools.

- Core components:
  - Common utilities that support the other Hadoop modules.
  - Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
  - YARN (Yet Another Resource Manager): A framework for job scheduling and cluster resource management.
  - MapReduce: A YARN-based system for parallel processing of large data sets.
- Add-ons and related projects:
  - Cluster/Job Management: Amari, ZooKeeper
  - Databases: Cassandra, HBase, Parquet
  - Streaming engines (for fast data applications): Flink, Kafka, Spark Streaming
  - Languages, libraries and compute engines: Pig, Hive, Mahout, Spark

Georgia Tech

The Hadoop Ecosystem



Georgia Tech

Installing Hadoop

- Single computer
- Cluster

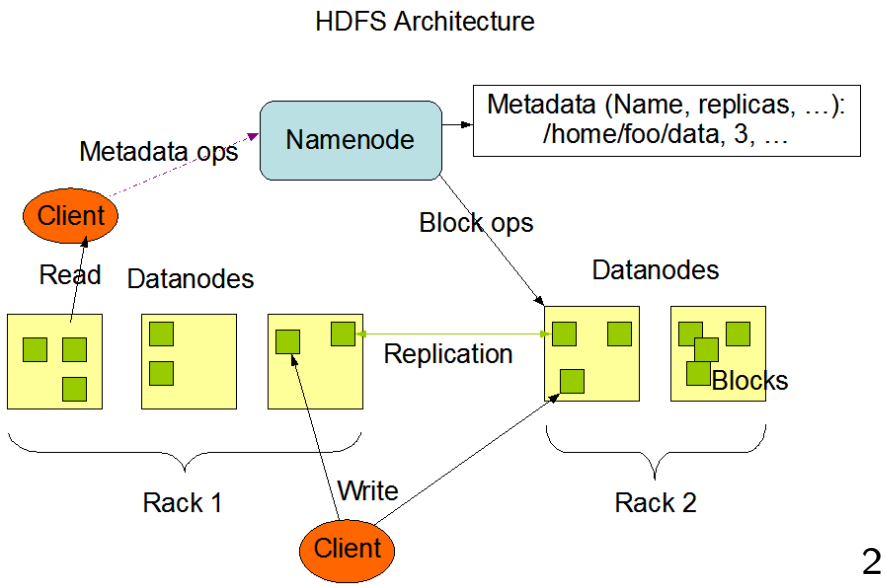
Georgia Tech

HDFS Assumptions and Goals

- Hardware Failures will happen. Detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.
- Streaming Data Access – high-throughput rather than interactive use. Trade a few POSIX requirements to increase data throughput.
- Large Data Sets – tens of millions of large files (gigabytes to terabytes each)
- Simple Coherency Model – write-once-read-many. After creation, files can only be appended to or truncated.
- "Moving Computation is Cheaper than Moving Data"
- Portability Across Heterogeneous Hardware and Software Platforms

Georgia Tech

HDFS Architecture



2

[2] http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

MapReduce

split - map - reduce

Example: Word Count

Canonical example.