

# **Data Processing in the Cloud**

**Parand Tony Darugar**

**<http://parand.com/say/>**

**[darugar@yahoo-inc.com](mailto:darugar@yahoo-inc.com)**

## What is Hadoop

Flexible infrastructure for large scale computation and data processing on a network of commodity hardware.

## Why?

A common infrastructure pattern extracted from building distributed systems

- Scale
- Incremental growth
- Cost
- Flexibility

## Built-in Resilience to Failure

- When dealing with large numbers of commodity servers, failure is a fact of life
- Assume failure, build protections and recovery into your architecture
  - Data level redundancy
  - Job/Task level monitoring and automated restart and re-allocation

## Current State of Hadoop Project

- Top level Apache Foundation project
- In production use at Yahoo, Facebook, Amazon, IBM, Fox, NY Times, Powerset, ...
- Large, active user base, mailing lists, user groups
- Very active development, strong development team

## Widely Adopted

- A valuable and reusable skill set
  - Taught at major universities
  - Easier to hire for
  - Easier to train on
  - Portable across projects, groups

## Plethora of Related Projects

- Pig
- Hive
- Hbase
- Cascading
- Hadoop on EC2
- JAQL , X-Trace, Happy, Mahout

## What is Hadoop

The Linux of distributed processing.

# **How Does Hadoop Work?**

## Hadoop File System

- A distributed file system for large data
  - Your data in triplicate
  - Built-in redundancy, resiliency to large scale failures
  - Intelligent distribution, striping across racks
  - Accommodates very large data sizes
  - On commodity hardware

## Programming Model: Map/Reduce

- Very simple programming model:
  - Map(anything)->key, value
  - *Sort, partition on key*
  - Reduce(key,value)->key, value
- No parallel processing / message passing semantics
- Programmable in Java or any other language (streaming)

## Processing Model

- Create or allocate a cluster
- Put data onto the file system:
  - Data is split into blocks, stored in triplicate across your cluster
- Run your job:
  - Your Map code is copied to the allocated nodes, preferring nodes that contain copies of your data
    - Move computation to data, not data to computation

## Processing Model

- Monitor workers, automatically restarting failed or slow tasks
- Gather output of Map, sort and partition on key
- Run Reduce tasks
  - Monitor workers, automatically restarting failed or slow tasks
- Results of your job are now available on the Hadoop file system

## Hadoop on the Grid

- Managed Hadoop clusters
- Shared resources
  - improved utilization
- Standard data sets, storage
- Shared, standardized operations management
- Hosted internally or externally (e.g., on EC2)

**YAHOO!**

# Usage Patterns

## ETL

- Put large data source (eg. Log files) onto the Hadoop File System
- Perform aggregations, transformations, normalizations on the data
- Load into RDBMS / data mart

## Reporting and Analytics

- Run canned and ad-hoc queries over large data
- Run analytics and data mining operations on large data
- Produce reports for end-user consumption or loading into data mart

## Data Processing Pipelines

- Multi-step pipelines for data processing
- Coordination, scheduling, data collection and publishing of feeds
- SLA carrying, regularly scheduled jobs

## Machine Learning & Graph Algorithms

- Traverse large graphs and data sets, building models and classifiers
- Implement machine learning algorithms over massive data sets

## General Back-End Processing

- Implement significant portions of back-end, batch oriented processing on the grid
- General computation framework
- Simplify back-end architecture

## What Next?

- Download Hadoop:
  - <http://hadoop.apache.org/>
- Try it on your laptop
- Try Pig
  - <http://hadoop.apache.org/pig/>
- Deploy to multiple boxes
- Try it on EC2