Apache Hama
a BSP for Advanced Analytics

Edward J. Yoon
@eddieyoon
edwardyoon@apache.org
I am ..

- Vice President of Apache Hama
- PMC member and Committer of ..
  - Apache BigTop
  - Apache MRQL
- Oracle corp. (2012 ~ 2013)
- Korea Telecom (2011 ~ 2012)
- NHN, corp. (2006 ~ 2011)
Agenda

1. Introduction of Hama
   a. What's Hama?
   b. Why Hama?
   c. Benchmarks

2. Use cases of Apache Hama
   a. Netflow Analytics at Korea Telecom
   b. SiteRank at Sogou.com

3. What's Next?
Introduction of Hama BSP
What's Hama?

- Apache Top Level Project
- Written in Java
- a general BSP computing engine
  - Java, Python, and C/C++ Interface
  - MRQL - BSP Query Language
- 10+ Active committers!
하마
Listed in Bossie Awards 2013

Apache Hama Project

Apache Hama is a pure BSP (Bulk Synchronous Parallel) computing framework on top of HDFS (Hadoop Distributed File System) for massive scientific computations such as matrix, graph and network algorithms.

Recent News

- June 25, 2013: release 0.6.2 available
- April 01, 2013: release 0.6.1 available
- November 28, 2012: release 0.6.0 available
- June 31, 2012: release 0.5.0 available
- May 17, 2012: Apache Hama graduated as a Top Level Project!

Why Hama and BSP?

Today, many practical data processing applications require a more flexible programming abstraction model that is compatible to run on highly scalable and massive data systems (e.g., HDFS, HBase, etc.). A message passing paradigm beyond Map-Reduce framework would increase its flexibility in its communication capability. Bulk Synchronous Parallel (BSP) model fits the bill appropriately. Some of its significant advantages over MapReduce and MPI are:

- Supports message passing paradigm style of application development
- Provides a flexible, simple, and easy-to-use, APIs
- Enables to perform better than MPI for communication-intensive applications
- Guarantees impossibility of deadlocks or collisions in the communication mechanisms
What's BSP?

- a Parallel computing model on **Message-Passing Architecture**
MapReduce vs. Hama BSP

Data-Intensive

Complex Computation-intensive
MPI on YARN vs. BSP

How can you solve the below problems of MPI?

- Data partitioning and data locality aware task scheduling
- Job Fault Tolerance
- Deadlock or Race Conditions
- Complexity of Interface

The BSP is answer!
Hadoop Ecosystem

HDFS

MapReduce

Apache Hama (BSP)

Apache MRQL

Spark, MPI, …, etc.

Hive, Mahout, …, etc.

YARN

HDFS
Why Hama?
Evolution of the WWW

● 1990 ~ : Web Documents

Web 2.0

Blog, Open API

Smartphone

Social Network

● ~ 2013 : Responsive Apps for multi-devices
Evolution of the Infrastructure

● 1990 ~ : Server/Web Hosting
  Google Apps
  Cloud Computing
  IaaS, PaaS, SaaS

● ~ 2013 : Cloud/App Hosting
Transition of the Technologies

● 2003 ~ :
  ○ SQL Database connectivity interface
  ○ Web-scale data processing
    ■ MapReduce
    ■ Hive, Pig, Mahout, ....

● 2007 ~ :
  ○ Key/Value interface
  ○ Realtime, ML and Graph Processing
    ■ Storm
    ■ Apache Hama!
    ■ GraphLab, Spark, …, etc.
So, why Hama?

- Simple and Flexible message-passing programming Interface

And,

- Machine Learning Package
  - K-Means clustering is almost 500x ~ 1000x faster than Mahout MR version

- Graph Package (Google's Pregel)
  - PageRank is almost 10 ~ 20x faster than MapReduce version
Benchmarks with 256 cores

- SSSP on random 1 Billion edges
  400 secs!

- PageRank on Wikipedia link DataSet, contains 5,716,808 pages and 130,160,392 links).
  17 secs!
Hama vs. Spark - KMeans
Use cases of Apache Hama
Netflow Analytics at Korea Telecom

Weather forecasting for Clouds

● 4 Full Racks
● Used as a Real-time event processing
  ○ Monitoring the network usage of each VMs as a real time
  ○ Detecting anomaly traffics
  ○ Sharing the risks among Servers
  ○ Billing, …, etc.
SiteRank at Sogou.com

Sogou.com runs SiteRank algorithm on a 7,200 cores Hama cluster.

- SiteRank is the ranking generated by applying the classical PageRank algorithm to the graph of Web sites.
- Dataset is about 400GB contains about 600M vertices and 6B edges
What's Next?

- Performance Improvement
- Develop the Hadoop Ecosystem
Hadoop Ecosystem

HDFS

MapReduce

Apache Hama (BSP)

YARN

Apache MRQL

Spark, MPI, …, etc.

Hive, Mahout, …, etc.
Thanks, Questions?