RED HAT :: CHICAGO :: 2009

SUMMIT

Birds of a Feather

Cloud Data Analytics

Tom Plunkett

Thomas.Plunkett@serenesoftware.com
Cloud Computing Architect
Serene Software
September 2, 2009

presented by

Birds of a Feather: Cloud Data Analytics

- Introduction
- Technology
- Use Cases
- Questions

Tom Plunkett

- Federal Government Experience
- IBM Certified SOA Solution Designer
- Patents
- Teach Cloud Computing and Java

NIST Cloud Computing Definition

- Essential Characteristics
  - On-demand Self Service
  - Ubiquitous Network Access
  - Location Independent Resource Pooling
  - Rapid Elasticity
  - Measured Service
- Delivery Models (SaaS, PaaS, IaaS)
- Deployment Models (Public, Private, Hybrid)

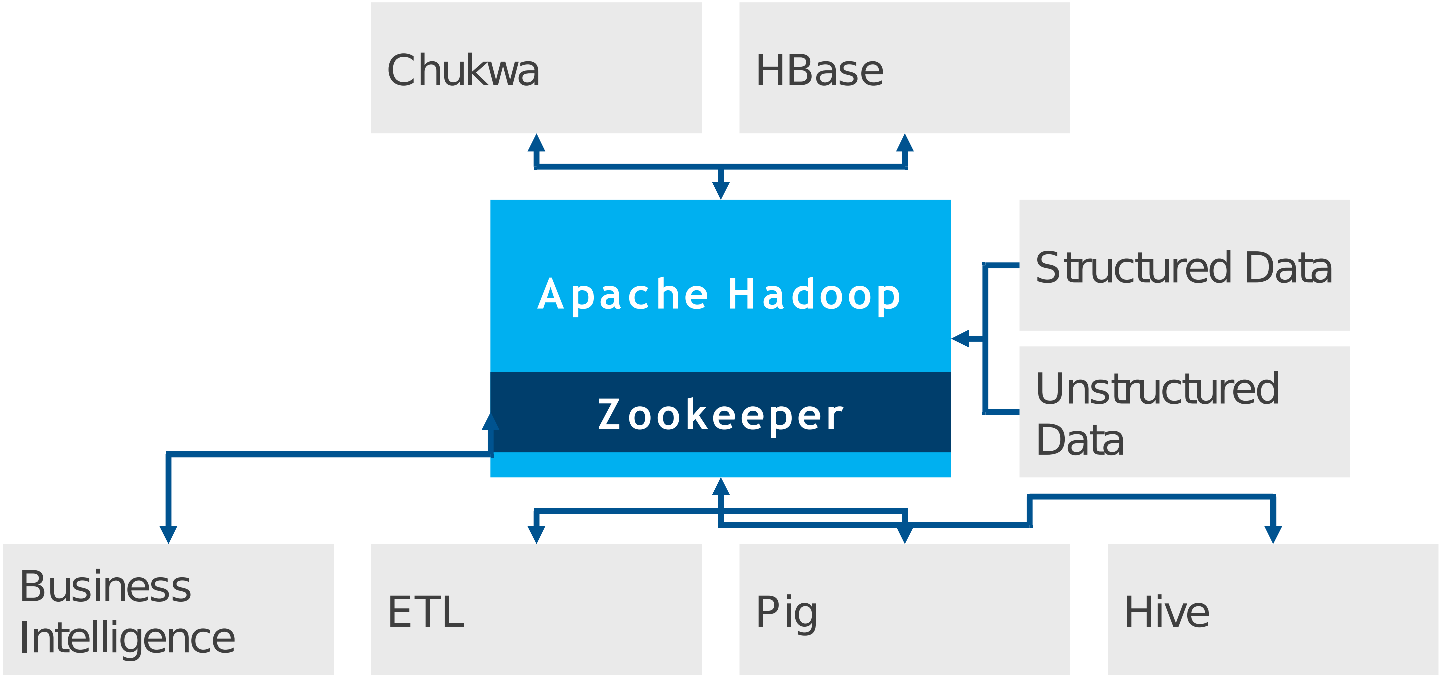
Google MapReduce

- Algorithm for computing distributed problems using a divide and conquer approach with a cluster of nodes
- Master node Maps input into smaller sub-problems and distributes the work to the cluster. A worker node may further map the work for a further cluster of nodes. The worker nodes then process the smaller problems, and return the answers back to the master node.
- Master node then Reduces the set of answers into the answer to the original problem

Apache Hadoop

- Open Source implementation of MapReduce
- Hadoop can store and process petabytes of data
- Many sub-projects and related projects
- Yahoo (more than 100k CPUs in more than 25k computers running Hadoop) and other companies make extensive use of Hadoop
- Hadoop has set TeraByte sort records

Apache Hadoop and Sub-projects



Apache Hadoop Sub Projects

- Chukwa -- Data collection system for monitoring and analyzing large distributed systems
- HBase -- similar to Google's BigTable, linear scalable distributed database for semi-structured data
- Hive -- Data warehouse infrastructure for large datasets with SQL-like query language
- Pig -- Functional language for data analysis
- Zookeeper --Configuration, naming, distributed synchronization, and group services

Use cases for Cloud Data Analysis

- Log Processing
- Event Detection
- Fraud Analysis
- Relationship Maps
- Relevance Ranking
- Trend Analysis
- Unstructured Data Analysis

Hadoop in Production

- Facebook: Producing summaries over large amounts of data to drive product decisions (page views, user growth, average time, performance numbers for ad campaigns, etc.)
- New York Times: Converting documents into pdf format
- Nutch Search Engine: Ranking and sorting URLs
- Rackspace: Log processing

Federal Use Cases

- DoD Intelligence Community Analysts
  - Petabytes of sensor data
- NASA Data Mining of Scientific Experiment Data
  - Petabytes of test data

Resources

http://hadoop.apache.org/

Books

Jason Venner, Pro Hadoop

Tom White, Hadoop: The Definitive Guide

Questions

Tom Plunkett

Cloud Computing Architect

(256)348-3667

Thomas.Plunkett@serenesoftware.com