Introduction to Big Data

Presented by: Jose Chinchilla

President, Agile Bay, Inc.

Jose Chinchilla

MCITP: Database Administrator, SQL Server 2008

MCITP: Business Intelligence SQL Server 2008

Current Positions:

President, Agile Bay, Inc.

President, Tampa Bay Business Intelligence User Group

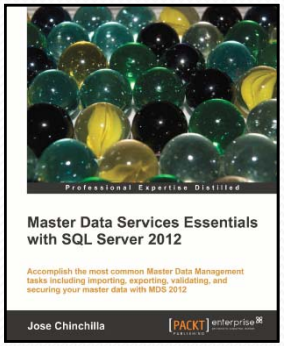
Regional Mentor, PASS Greater Southeast

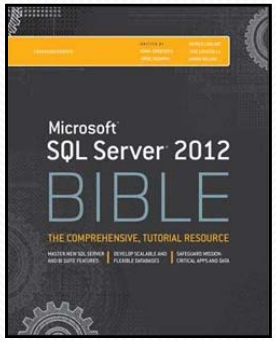
Blog: http://www.sqljoe.com

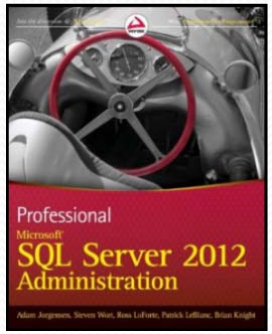
Twitter: @sqljoe

Linked-in: http://www.linkedin.com/in/josechinchilla

Email: jchinchilla@sqljoe.com







Customers & Partners

agilebay

www.agilebay.com

Data Warehouse Analytics Master Data Training

Session Agenda

- What is Big Data?
- What is Hadoop?
- BI vs. Hadoop
- The Hadoop Ecosystem
- Real-world Use Cases
- Demo:

Terms and Acronyms

- Hadoop:
  Apache project (open source) project to develop software for reliable, scalable, distributed computing.
- Cluster:
  A group of computers (nodes) linked together to perform a highly-available and high computation work
- HDFS
  distributed file system that provides high-throughput access to application data.
- YARN
  A framework for job scheduling and cluster resource management.
- MapReduce

A YARN-based system for parallel processing of large data sets.

What is Big Data?

What is Big Data?

- Big data is a **buzzword**, used to describe a **massive volume** of both structured and unstructured data that is so large that it's **difficult to process using traditional database and software techniques**.
- Solves challenges related to:
  - Capture
  - Maintenance
  - Storage
  - Search
  - Share
  - Transfer
  - Analysis
  - Unstructured
  - Visualization
  - Scale
  - Availability
  - Recovery
- Volume, Velocity, Variety

What is Hadoop?

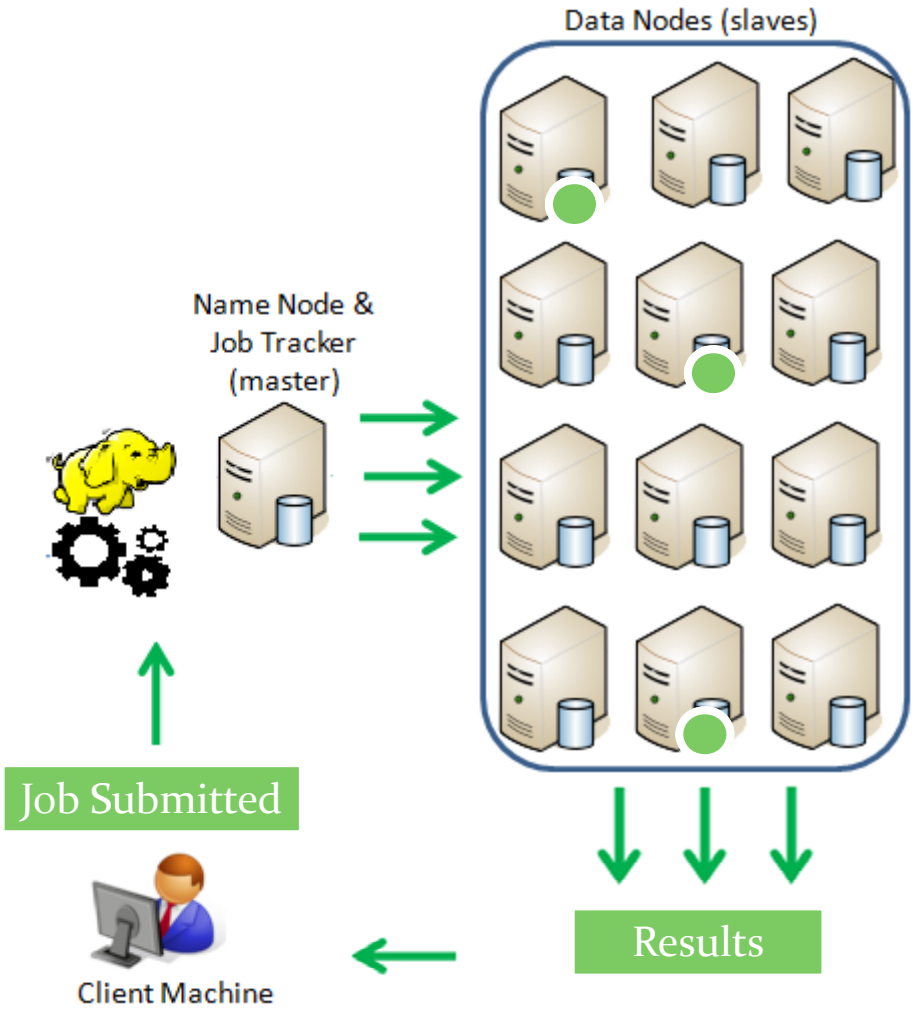
What is Hadoop

- Apache open source project
- Distributed computing (cluster)
- Ecosystem

  - Ambari
  - HBase
  - Avro
  - Cassandra
  - Chukwa
  - Hive
  - Mahout
  - Pig
  - ZooKeeper

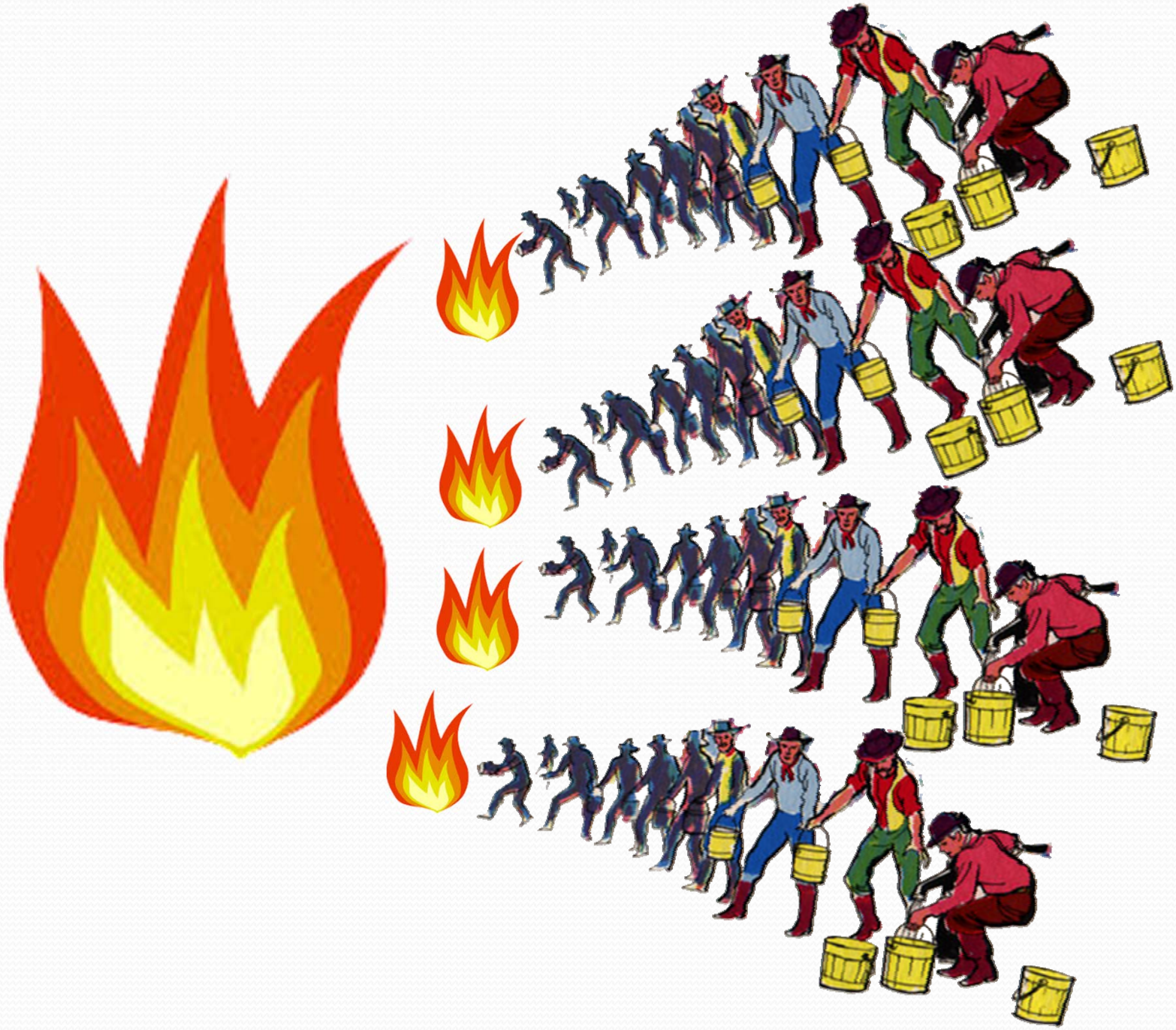
Distributed Computing & MapReduce

MapReduce



Map Phase

Reduce Phase



BI vs. Hadoop?

BI vs. Hadoop

- Hadoop not a replacement of BI
- Extends BI capabilities
- BI = Scale up to 100s of Gigabytes
- Hadoop = From 100s of Gygabytes to Terabytes (1,000s og Gygabytes) and Terabytes (1,000,000 Gigabytes)

Thank you for attending!

Q & A

Blog: www.sqljoe.com

Twitter: @sqljoe

Linked-in: http://www.linkedin.com/in/josechinchilla

Email: jchinchilla@agilebay.com