INL HPC Big Data Analytics

Big Data Conference May 19, 2015

Shad Staples – HPC Big Data Lead
Scott Jeffery – HPC Big Data
Wayne Simpson – Enterprise Innovation Architect

www.inl.gov

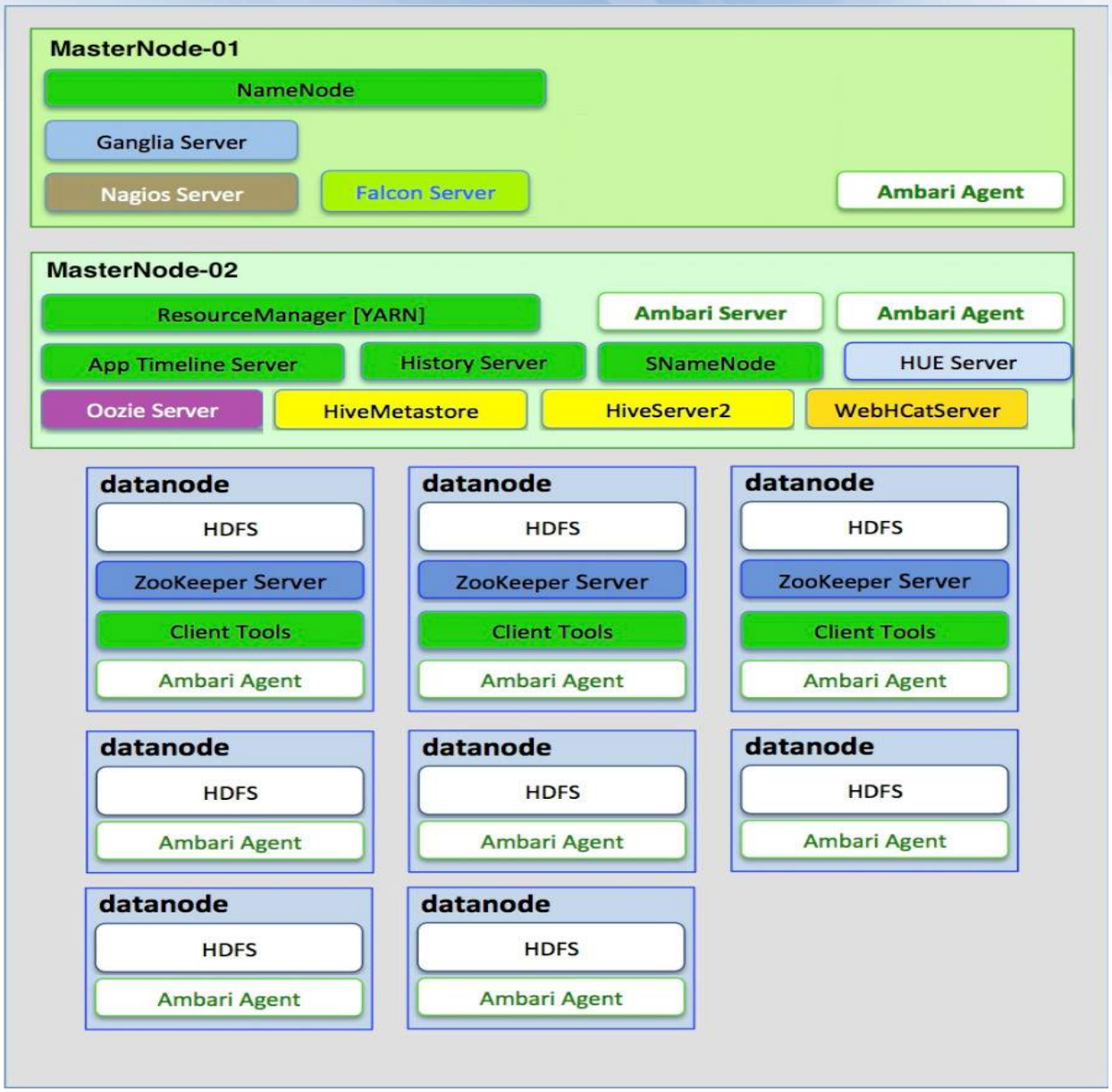
Idaho National Laboratory

History

- INL IM Initiative – HPC Big Data Analysis
  - Enable researchers and scientist
  - Provide a central location vs silo models
  - Provide a service of Volume, Variety, and Velocity
- Platform
  - Cassandra vs **Hadoop**
  - Cloudera vs **Hortonworks**
- Hardware – HP vs Dell
  - Ten nodes two masters and eight slaves
  - One terabyte RAM
  - 350 terabytes spinning storage
  - 320 cpu cores

Challenges & What Worked

- Internal Infrastructure
  - Firewall
  - Proxy
  - Policies
- Location – Internal vs Public Enclave
- Learning Curve – Drinking from a Firehose
  - Installation of a Hadoop Cluster - Ambari
  - Components – Hive, Pig, Sqoop, HDFS, etc.
  - Manual User Management
- Hortonworks Documentation & Support
- ESRI GIS Github Tools - http://esri.github.io/gis-tools-for-hadoop/
- Tool Set - R Studio, Python, Java, QGIS, ArcMap


MasterNode-01
NameNode
Ganglia Server
Nagios Server
Falcon Server
Ambari Agent
MasterNode-02
ResourceManager [YARN]
Ambari Server
Ambari Agent
App Timeline Server
History Server
SNameNode
HUE Server
Oozie Server
HiveMetastore
HiveServer2
WebHCatServer
datanode
HDFS
ZooKeeper Server
Client Tools
Ambari Agent
datanode
HDFS
ZooKeeper Server
Client Tools
Ambari Agent
datanode
HDFS
ZooKeeper Server
Client Tools
Ambari Agent
datanode
HDFS
Ambari Agent
datanode
HDFS
Ambari Agent
datanode
HDFS
Ambari Agent
datanode
HDFS
Ambari Agent
datanode
HDFS
Ambari Agent

Accomplishments

- Open NLP & Solr/Nutch
  - Natural Language Processing & Information Extraction – Ryan Hruska Presentation
- GIS Data Processing
  - Geographical Data Processing - Sera White Presentation
  - Data Layering – Visualization
  - Energy Data Analytics
- R & Statistical Data Analytics
  - Department of Homeland Security – Critical Infrastructure Assessment - Data Analysis

What's Next?

- Grow the Cluster - Six additional nodes one master and five slaves
- Hadoop on the HPC Cluster
- Administrative
  - User Management – Apache Ranger
  - Tuning – Can we get all the CPU's working on a problem?
  - Automated Cluster Provisioning - PXE Boot / Werwulf + Ambari?
- Data Curation & Centralization of Data Sources
  - Metadata Standardization
  - Dublin Core
  - User Searchable
- Solutions
  - Internal Document Enterprise Search
  - Cyber Security Analytics
    - Sqrrl
    - Spark

Idaho National Laboratory