SFO15-TR6: Hadoop on ARM

Nachiket Bhoyar
Steve Capper

Presented by
Nachiket Bhoyar
Steve Capper

Date
Wednesday 23 September 2015

Event
SFO15
Agenda

1. Quick intro to Hadoop stack.
2. Summary of our work.
3. Demo time!
4. Q & A
The Hadoop Stack

And lots more components!.....
The Hadoop Distribution

- **LOTS** of components fit with Hadoop.
- *Hadoop distros* package these.
- The Open Data Platform Initiative has just been formed to promote compatibility between Hadoop distros.
Our Hadoop work

- Open Data Platform is in early days.
- A Hadoop distro was needed for us to start experimenting with for AArch64.
- We chose to start with Hortonworks (who are a member of Open Data Platform).
- We will move on to work with Open Data Platform distributions.
AArch64 Hadoop Work

- A lot of ramp up on build systems (Ant, Ivy, Maven, Gradle…), and tweaking build logic.
- We had to stop builds downloading the x86 version of node.js then running it on ARM…
  - io.js was needed as it worked with AArch64 V8 JS.
- Otherwise, things mostly just worked.
- Upstream Hadoop and Spark are being investigated too.
OpenJDK Work

- Building and testing Hadoop + Spark has given the AArch64 OpenJDK a very good stress test.
- A bug has been found and it has been fixed in the 1508 OpenJDK release:
  - [https://bugs.openjdk.java.net/browse/JDK-8133842](https://bugs.openjdk.java.net/browse/JDK-8133842)
Future work

● We need to package up everything:
  ○ currently tricky as we don’t have the deb/rpm logic,
  ○ some build systems appear to download the internet
  ○ which is very bad in areas with no local mirrors!

● Clusters to be deployed + tested + profiled.

● Workloads that are representative of real world need to formulated and executed as well as micro-benchmarks.
Demo Time!
Useful H$_2$O Links

- **H$_2$O**: [http://h2o.ai/product/](http://h2o.ai/product/)
- **Downloads**: [http://h2o.ai/download/](http://h2o.ai/download/)
- **HDP + H$_2$O video**: [https://www.youtube.com/watch?v=KigG7rPBNHM](https://www.youtube.com/watch?v=KigG7rPBNHM)
- **H2O Airlines Demo Video**: [https://www.youtube.com/watch?v=bInMSgZhDd4](https://www.youtube.com/watch?v=bInMSgZhDd4)
Thank you for your attention!

Any questions/comments?
Backup Slides
Agenda

1. What is H$_2$O?
2. What is a Flow?
3. H$_2$O with Hadoop
4. System Configuration
5. Demo
6. Summary
What is H₂O?

- Data collection is easy. Decision making is hard.
- H₂O derives insight using faster and better predictive modelling.
- Combines power of:
  - Highly advanced algorithms
  - Freedom of open source
  - Capacity of scalable in-memory processing
- Processes big data on single or multiple nodes.
- Supports R, Python, Scala, Java and ReST API.
- Easy integration with Hadoop
H₂O Stack
What is a Flow?

- A Flow is an open-source user interface for H₂O
- Allows user to combine code execution, text, mathematics, graphs, and rich media in a single document
- In simplest sense, it’s a sequence of executable cells
- Cells can be modified, rearranged or saved to library
- Each cell has input field to:
  - Enter commands
  - Define functions
  - Call other functions
  - Access other cells/objects in the flow
H$_2$O with Hadoop

- H$_2$O can be run as an application in Hadoop
- It is run as a mapper process on each node
- Easy integration of data from HDFS
- Shows Cluster Status:
  - GC status, Disk usage, System usage, System load, etc.
  - Water meter to show status of cores
System Configuration

- Cluster - 6 nodes of AMD Opteron A1100 ARM64 servers
- Memory - 64GB per node
- OS - Fedora 22
- JDK - Linaro Open JDK 1.8 15/08 release
- Hadoop - Hortonworks HDP 2.6.0-SNAPSHOT
- H2O version - h2o-3.0.0.30-hdp2.2
Model Building Scaling

- Linear scaling observed for both 32GB and 64GB
File Parsing Scaling

- This phase is network dependent
- A linear scaling observed for 10GigE
- Network bottleneck observed for 1GigE going beyond 2 nodes
Summary

- AMD Opteron A1100 and Linaro Open JDK 1.8 scale linearly w.r.t. number of nodes on H2O
- 10GigE ethernet scales linearly whereas 1GigE suffers from bottleneck
Summary - $H_2O$

- $H_2O$ helps to easily apply math and predictive analytics to solve challenging business problems
- With $H_2O$, you can:
  - Make better predictions using ready-to-use algorithms and processing power to analyze: bigger data sets, more models and more variables
  - Work with your existing languages and tools
  - Extend the platform seamlessly into your Hadoop environments
- It is Open Source
Summary - Flow

- Import data Files > Build Models > Iteratively Improve them > Make predictions
- Easy-to-use Modern Graphical Interactive WebUI
- Access any H₂O object in well-organized tabular data