The Big Data Puzzle - Where Does the Eclipse Piece Fit?

March 9, 2016
About me:


Developing applications and frameworks based on Eclipse technology since 2009.

Specialize in “Horizontal Integration” - combining existing technology to provide new capabilities.

Find me online:

Email – jlangley@cohesionforce.com
Twitter – @jperiodlangley
Google+ JLangley
About CohesionForce:

CohesionForce Inc. (CFI) is a Technology and Services Veteran Owned Small Business headquartered in Huntsville, Alabama. CFI has progressed from our founding in 1998 with measured growth focused on providing innovative solutions and collaborative teaming for complete support to our Prime and Government customers.

Core Values:

- Better tools are needed to solve the problems of tomorrow
- Disciplined engineering should be easy
- Systems should be fulfilling to use
Summary:

The tools used for Big Data analytics seem to be converging on the Apache Software Foundation as a home. As an Eclipse and Apache integrator, CohesionForce has found a useful fit for Eclipse projects when used as tooling for an underlying Apache project runtime.

Using file formats such as Apache Avro and Parquet allows us to query the data using a SQL “like” language with Apache Hive and Spark.
Scope:

Big Data is well... a big topic. In this talk, we will focus on the following:

1. Sample data used to approximate the problem space
2. Apache Projects that CohesionForce has used for data analysis
3. Eclipse Projects that CohesionForce has used/developed for tooling
4. Thoughts on future work with Eclipse for Data Science
Sample Data:

A sizable list of data available for use can be found here: https://aws.amazon.com/public-data-sets/

These data sets are a mix of textual, spatial, image, and video. They closely approximate the size of the data that we needed, but did not have the combination of factors that we were looking for.

We needed data for events that have an ID, type, location, and a time. We were able to create a suitable data set by transforming taxi data available here: http://www.andresmh.com/nyctaxitrips/

After converting the data to a common format (DIS EntityStatePDU), the result was just over 174M events.
Distributed Interactive Simulation:

Distributed Interactive Simulation (DIS)

DIS is an IEEE standard (IEEE-1278.1) developed by the Simulation Interoperability Standards (SISO) and approved by IEEE.

The EntityStatePDU contains fields such as

- EntityID – Site, Application, Entity
- EntityType – 7-tuple enumeration
- EntityLocation – x/y/z Geocentric WGS-84
- EntityOrientation – psi/theta/phi Euler angles
- EntityVelocity – x/y/z along the orientation axis
### Apache Big Data Projects:

<table>
<thead>
<tr>
<th>Apache Ambari</th>
<th>Apache Helix</th>
<th>Apache Tajo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache Avro</td>
<td>Apache Ignite</td>
<td>Apache Tez</td>
</tr>
<tr>
<td>Apache Bigtop</td>
<td>Apache Kafka</td>
<td>Apache VXQuery</td>
</tr>
<tr>
<td>Apache BookKeeper</td>
<td>Apache Knox</td>
<td></td>
</tr>
<tr>
<td>Apache CouchDB</td>
<td>Apache MetaModel</td>
<td></td>
</tr>
<tr>
<td>Apache Crunch</td>
<td>Apache ORC</td>
<td></td>
</tr>
<tr>
<td>Apache DataFu</td>
<td>Apache Oozie</td>
<td></td>
</tr>
<tr>
<td>Apache DirectMemory</td>
<td>Apache Parquet</td>
<td></td>
</tr>
<tr>
<td>Apache Drill</td>
<td>Apache Phoenix</td>
<td></td>
</tr>
<tr>
<td>Apache Falcon</td>
<td>Apache REEF</td>
<td></td>
</tr>
<tr>
<td>Apache Flink</td>
<td>Apache Samza</td>
<td></td>
</tr>
<tr>
<td>Apache Flume</td>
<td>Apache Spark</td>
<td></td>
</tr>
<tr>
<td>Apache Giraph</td>
<td>Apache Sqoop</td>
<td></td>
</tr>
<tr>
<td>Apache Hama</td>
<td>Apache Storm</td>
<td></td>
</tr>
</tbody>
</table>
CohesionForce experience with:

- Apache Hadoop
- Apache Hive
- Apache Avro
- Apache Spark
- Apache Parquet
Initial Implementation:

- Eclipse BIRT
- Apache Hive
- Apache Yarn
- Hadoop NN
- Hadoop DN
- Hadoop DN
- Hadoop DN
Second Implementation:

- Eclipse BIRT
  - Apache Hive
  - Spark Master
  - Apache Yarn
  - Hadoop NN
    - Spark Worker
      - Hadoop DN
    - Spark Worker
      - Hadoop DN
    - Spark Worker
      - Hadoop DN
Third Implementation:

- Eclipse BIRT
  - Apache Hive
    - Spark Master
    - Hadoop NN
      - Spark Worker
        - Hadoop DN
      - Spark Worker
        - Hadoop DN
      - Spark Worker
        - Hadoop DN
Fourth Implementation:

- Spark Shell
  - Spark Master
    - Hadoop NN
      - Spark Worker
        - Hadoop DN
      - Spark Worker
        - Hadoop DN
      - Spark Worker
        - Hadoop DN
Current Implementation:

Browser

Apache Zeppelin

Spark Master

Hadoop NN

Spark Worker

Hadoop DN

Spark Worker

Hadoop DN

Spark Worker

Hadoop DN
Don't forget to show Zeppelin!

See other attachment on session
Workflow:

1. Log Data
2. Load data into Hadoop
3. Execute Hive Queries
4. Generate Logger
5. Generate Queries
6. Data Model
7. Export Results

Entry → Hadoop-Hive
Exit
Eclipse Tooling:

Eclipse-Avro: The purpose of this repository is to provide the capability to store EMF data files in the Apache Avro format. The Acceleo project is used to generate an Avro Schema based on a given EMF schema along with an AvroResourceImpl that can be used in the place of the XMIResourceImpl to load and save data using the common EMF methodology. https://github.com/LangleyStudios/eclipse-avro

DIS Toolkit: The DIS Toolkit provides an EMF model based on the DIS schema provided by the OpenDIS codebase, along with a generated data logger that stores files in compressed binary using the Apache Avro format. https://github.com/CohesionForce/dis-toolkit

AvroToParquet: a simple command line converter for Apache Avro to Apache Parquet file formats. https://github.com/CohesionForce/avroToParquet
Other Eclipse Based Projects:

**Eclipse BIRT** - Visualizing Big Data with Hadoop and BIRT

**Talend Open Data Studio** – Start working with Hadoop and NoSQL databases today using simple, graphical tools and wizards to generate native code that leverages the full power of Hadoop.
https://www.talend.com/download/talend-open-studio

**DataStax DevCenter** - a free, Eclipse-based tool, which is designed to be a lightweight visual interface that provides developers and others with the ability to easily run CQL queries against Cassandra, view query results, and perform other related tasks.
http://www.datastax.com/what-we-offer/products-services/devcenter
More Eclipse Based Projects:

Architect - provides an Eclipse-based workbench in which data scientists can get their job done, in other words, an integrated development environment (IDE) targeted specifically at data scientists.
http://www.openanalytics.eu/products

StatED - an Eclipse based IDE (integrated development environment) for R. It offers a set of mature tools for R coding and package building. This includes a fully integrated R Console, Object Browser and R Help System, whereas multiple local and remote installations of R are supported.
http://www.walware.de/goto/statet
https://github.com/walware/statet
Thoughts on Future Direction:

Eclipse Sirius could be used to create graphical editors or visualizations of component deployment, schemas, queries, etc.

XText could be used to provide editors for Hive Query Language, while also providing binding to types retrieved from data schemas. This would allow a user to write queries with syntax highlighting, code completion, and validation before execution.

XTend could be used to generate loggers, transforms, and other tasks based on a data model. The config files needed for Hadoop & Spark could be generated based on a modeled laydown.

Provisioning bundles, starting containers.
Interested?

We would like to compile a working list of ideas on this subject.

We would also like to identify potential users of these types of tools to be sure that we implement the proper feature set.

If you need help with integrating any of the tools/concepts that have been covered, please:

Contact us:

CohesionForce
Phone: 256-562-0600
Email: jlangley@CohesionForce.com
Evaluate This Session

1. Sign-in: [www.eclipsecon.org](http://www.eclipsecon.org)

2. Select session from schedule

3. Evaluate: [+1] [0] [-1]