DA 320 Assignment 8 M. Blanco

**Part 1: Research Analysis**

There is not wrong or right answer when selecting of the following technologies. Many co-exist and compliment each other. Some best suited for certain applications.

**Problem 1:** Do a side-by-side comparison of Apache Pig and the following technologies in regards to writing Hadoop applications. Make sure you include the advantages and disadvantages of each, as well as when to use each technology over the other.

|  |  |  |
| --- | --- | --- |
|  | Pig | Hive |
| Disadvantage | * Pig Latin is not a full programming language, only a data flow language * Pig is SQL like but varies to a great extent * Operates on the client side of a cluster | * Declarative (SQL Dialect) * Directly leverages SQL and is easy to learn for database experts * Hive only understands a few file format * Operates on the server side of a cluster |
| Advantage | * Pig is meant for Programming * Pig Hadoop is very easy to learn read and write if you are familiar with SQL * Pig is usually more efficient than Hive as it has many high-quality codes * Pig consistently outperformed Hive for most of the operations except for grouping of data. * Part of Hadoop ecosystem | * Hive is meant for creating reports * Part of Hadoop ecosystem |

Both Hive Hadoop and Pig Hadoop will help to achieve the same goals.

|  |  |  |
| --- | --- | --- |
|  | Pig | Cascading |
| Advantage | * Scripting language * Static and dynamic typing | * Cascading is neither a scripting nor a SQL-oriented language—it is a set of .jars that define data processing APIs, integration APIs, as well as a process planner and scheduler * Weak typing |
| Disadvantage | * Manual save | * Run slower than native Hadoop * Ability to re-run a   flow automatically  from the last saved  checkpoint |

|  |  |  |
| --- | --- | --- |
|  | Pig | Java |
| Disadvantage | * Scripting language * Pig Latin code can be extended through various user defined functions that are written in Java | * High level programming language * Gives developer more options to write User Defined functions |
| Advantage | * Requires less lines of code to implement MapReduce than Java * Pig runs on top of MapReduce and abstracts Java MapReduce jobs away from developers. * The Pig Latin script is easier to read for someone without a Java background. | * Requires more lines of code to implement MapReduce than Pig * Higher learning curve |

Pig is preferred over Java for its simplicity.

|  |  |  |
| --- | --- | --- |
|  | Pig | MapReduce |
| Disadvantage | * Scripting language * Code efficiency is relative less | * Compiled language * Code efficiency is high * Expects the programming language skills for writing the business logic |
| Advantage | * Higher level of abstraction * Less lines of code * Development effort is relative less * PIG is Data flow language | * Lower level of abstraction * More lines of code * Development effort is more involved * MapReduce is Data Processing Paradigm |

Pig program is executed internally it converts into a MapReduce job and process the data.

References:

<https://www.pluralsight.com/blog/tutorials/pig-vs-java-mapreduce>

<https://www.linkedin.com/pulse/technical-differences-between-pig-hadoop-hive-jims-lajpat-nagar/>

<https://stackoverflow.com/questions/14052796/when-to-use-pig-and-when-to-use-java-for-mapreduce>

<https://hortonworks.com/blog/cascading-hadoop-big-data-whatever/>

**Problem 2:** Do a side-by-side comparison of Cascading and the following technologies in regards to writing Hadoop applications. Make sure you include the advantages and disadvantages of each, as well as when to use each technology over the other.

Cascading is a collection of applications, languages, and APIs for developing data-intensive applications.

Apache Storm is a free and open source distributed real time computation system.

|  |  |  |
| --- | --- | --- |
|  | Cascading | Storm |
| Advantage | Hides the underlying complexity of MapReduce jobs  Major open source project | * Storm works on individual events * Reliably process unbounded streams of data, doing for real time processing what Hadoop did for batch processing |

Use storm when the focus is on stream processing and CEP-style processing.

Cascalog is an extension to Cascading that enables application development with Clojure.

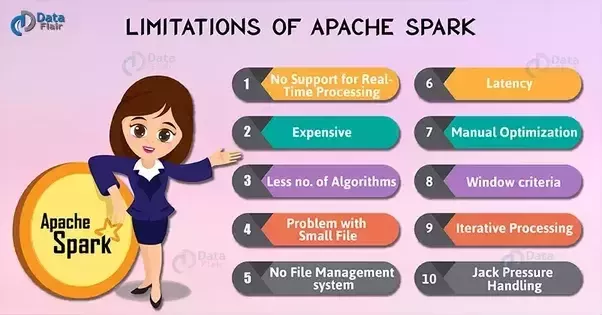
|  |  |  |
| --- | --- | --- |
|  | Cascading | Cascalog |
| Advantage | Hides the underlying complexity of MapReduce jobs  Major open source project | * Develop and test queries right inside of the Clojure REPL * develop processing workflows with extreme speed * Take advantage of existing libraries with very little effort, leading to much shorter development cycles |

Apache Spark™ is a fast and general engine for large-scale data processing.

|  |  |  |
| --- | --- | --- |
|  | Cascading | Spark |
| Disadvantage |  | * Consumes a lot of Memory * Requires large resources |
| Advantage | * Wrapper over Hadoop MapReduce | * Complements MapReduce for analytics workload * Spark utilize a rich set of Cascading extensions * SQL, streaming, and complex analytics * High-level language operators for streaming data * Fault-tolerant semantics * Support for merging streaming data with historical data   Write applications quickly in Java, Scala, Python, R.  In-memory distributed computation framework |
|  | * Popular open source Java-based framework for building data pipelines in Hadoop | * In-memory data processing framework * works on micro-batches |

Spark is best if an interactive shell for data exploration using API calls is required.





<https://blog.cloudera.com/blog/2013/11/cascading-spring-and-spark-development-choices-for-cdh-users-expand/>

<http://www.hadooptpoint.org/advantages-and-disadvantages-of-apache-spark/>

<https://scalding.io/2014/10/running-scalding-on-apache-spark/>

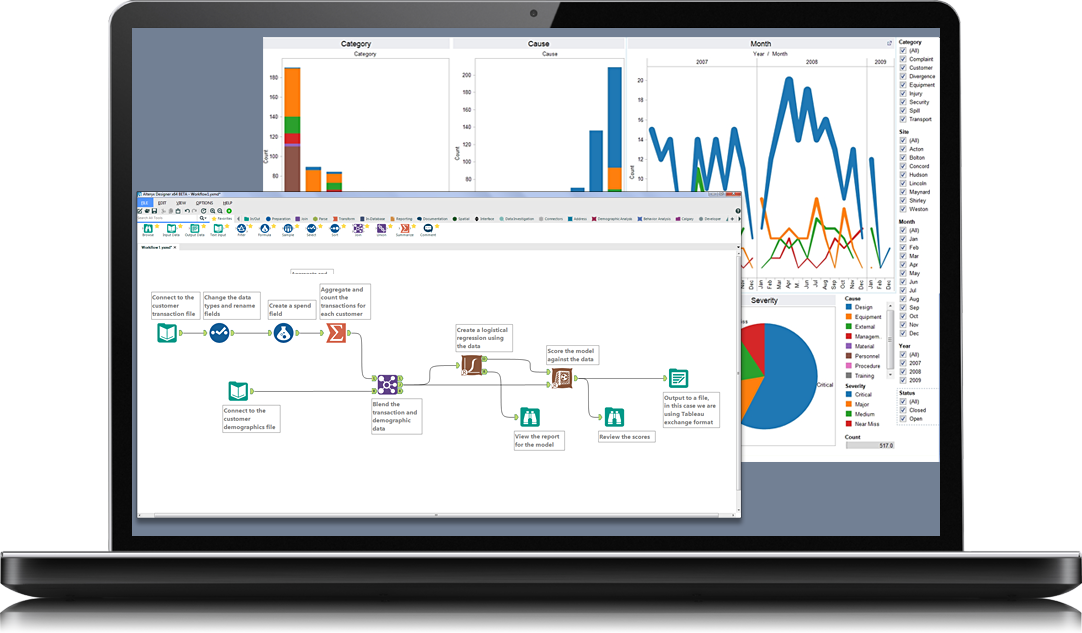
<https://www.quora.com/What-are-the-pros-and-cons-of-Apache-Spark>

**https://www.infoworld.com/article/2854894/application-development/spark-and-storm-for-real-time-computation.html**

**Problem 3:** Research data workflow software packages, at least one commercial and at least one open-source. Do a side-by-side comparison emphasizing advantages and disadvantages of each, as well as when to use each technology over the other.

### Alteryx

Alteryx is a self-service data analytics platform that will enable team to prep, blend, and analyze data, then deploy and share analytics.

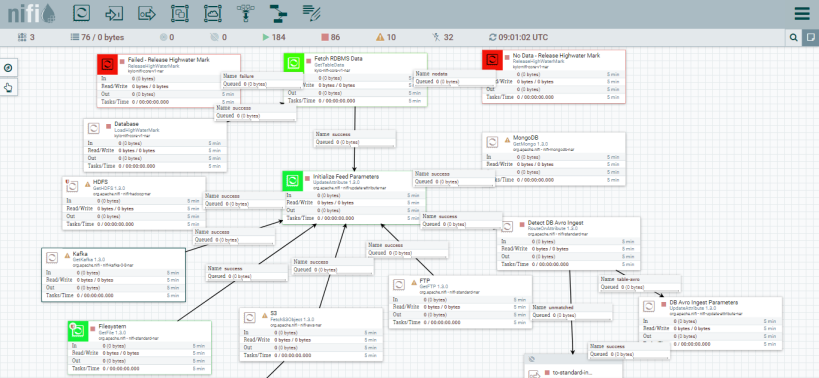


Source 1 Alteryx

|  |  |
| --- | --- |
| Alteryx Advantages | |
| Repeatable workflow | Automate manual data tasks |
| No coding | Drag and drop, intuitive user interface |
| Data volume & variety | Process data from various data source |
| Analytics at scale | Customize for easy consumption |
| Timely | Deeper insights in hours |
| Advanced analytics | Predictive, statistical, and spatial analytics |
| Learning | Docs, tutorial videos and Alteryx community |
| Alteryx Disadvantages | |
| Cost | Cost per year  The Alteryx Designer costs $5,194  Alteryx Server costs $58,500  Analytics Gallery costs $1,950 |
| Joins | Difficulty to do a true left to right join when data blending |
| Map visualization | Not the best tool for outputting presentation worthy maps |
| Run time | Disappointing run times when processing billion of records |

### Kylo

Kylo is an open source enterprise-ready data lake management software platform for self-service data ingest and data preparation with integrated metadata management, governance, and security.



|  |  |
| --- | --- |
| Kylo Advantages | |
| Data Ingestion | Self-service data cleansing, validation, and automatic profiling  Pipeline template mechanism |
| Data Preparation | Leverages the latest capabilities of Apache Spark |
| Data Discovery | Integrated metadata repository and key capabilities for data exploration |
| Data Monitoring | Integrated metadata repository and key capabilities for data exploration |
| Design | Design batch or streaming pipeline templates |
| Learning | Docs, tutorial videos |
| Kylo Disadvantages | |
| Beta | It was launched in late 2017. Beta version. |
| Installation | Requires a number of tools to be pre-installed |

References:

<http://www.zdnet.com/article/teradata-open-sources-kylo-data-lake-management-software/>

<https://www.thinkbiganalytics.com/wp-content/uploads/2017/01/FAQ_Kylo_v2.pdf>

http://www.treselle.com/blog/kylo-setup-for-data-lake-management/

**Part 2:**

**Problem 4:** Recall the scenario in Project 5 regarding BV designs [(http://bvdesigns.us/about.html)](http://bvdesigns.us/about.html), the custom jewelry maker have the students attempt to list relevant data stores, data flows, processes, and sources and sinks. Draw both a context diagram and a level-1 diagram that could represent the selling system. Note: since we are not able to observe sales transactions, you may make reasonable assumptions, and make sure you list your assumptions clearly.



