IKANOW

HOW INFINIT.E USES HADOOP TO ENABLE LARGE SCALE DOCUMENT ANALYSIS

By Craig Vitter, IKANOW Professional Services Engineer

It's hard to understate the impact Apache Hadoop has had on the ability of organizations to analyze truly massive data sets. Traditional data analysis tools have been designed around very efficient algorithms functioning in a linear fashion. As the volume of data to be analyzed grew exponentially, these algorithms are constrained by the ability of organizations to scale their hardware vertically via additional processors, memory, and faster storage devices. Ultimately most true big data problems have surpassed the capability of traditional data analysis tools and algorithms to function in a cost effective manner leaving fast swathes of data un-mined.

The introduction of the Open Source Hadoop platform and the MapReduce parallel data processing paradigm is a game changer for organizations unable to surmount the barriers erected by big data problems using traditional tools, for the first time, organizations have access to a low barrier-to-entry platform that allows them to take advantage of commodity hardware to build inexpensive data processing clusters without the traditional linear data analysis constraints.

Note: *If you are new to Hadoop and MapReduce IKANOW recommends visiting hadoop.apache.org to learn more.*

Of course simply installing a Hadoop cluster doesn't magically solve all of an organization's big data problems. For example, in order to realize value from big data analysis, organizations need to quantify the problems they are trying to solve and collect the data they will need (in a format accessible to their analysis tools) to actually solve those problems. Managing an elaborate infrastructure to feed a platform like Hadoop with data from a large variety of disparate sources can be a daunting challenge, one that keeps many organizations from utilizing the full power of their Hadoop clusters.

This white paper introduces the reader to IKANOW's Infinit.e Open Source platform and how its features help organizations unify their data and MapReduce jobs in a manner that accelerates the adoption of Hadoop.

THE INFINIT.E-HADOOP INTEGRATION

IKANOW's Infinit.e is an Open Source knowledge discovery and analysis platform that empowers deep analysis on structured and unstructured documents (database records, XML documents, RSS news feeds, Tweets, PDFs, etc.). The Infinit.e platform merges these disparate data types into one unified semi-structured format and exposes this data to discovery and analysis via a REST based API.

Although the Infinit.e platform is fully capable of operating as a stand-alone big data analysis tool, IKANOW's engineers architected it specifically to support rich levels of interaction with the full suite of open source data tools available including Hadoop. Infinit.e has the following components that support seamless integration with Hadoop:

MapReduce Plugin Manager
The Infinit.e MapReduce Plugin Manager is a web-based interface that allows developers to upload their MapReduce Java JARs to Infinit.e (via the REST based interface), configure the MapReduce jobs, and schedule the jobs to run on an ad-hoc or scheduled basis.

Infinite API
Infinit.e's REST based API is the primary means of interacting with Infinit.e and is used to manage MapReduce jobs (Java JARs and their configuration information) and execute MapReduce jobs via the Infinit.e platform.

Infinit.e Core Server
The Infinit.e Core Server is responsible for managing the scheduled processes that are central to the platform's functionality including harvesting source data and executing scheduled Hadoop MapReduce jobs.

MongoDB Hadoop Connector
The MongoDB Hadoop Connector is an Open Source plugin for Hadoop that allows MapReduce jobs to use MongoDB databases as both an input and output source.

These components combine to provide a number of beneficial features to organizations looking to adopt Hadoop, including the ability to:

Read From/Write to One Central Data Store
Using Infinit.e as a central data store, with its unified document format, greatly reduces the time and effort required to make disparate data sources available for Hadoop to process and to store the results of your MapReduce jobs.

Inject queries and configuration into a plugin at run time
The Infinit.e MapReduce plugin architecture allows developers to write generic plugins that accept MongoDB queries at run time making it easier to reuse and share plugins across multiple problem sets. Additionally, each MapReduce job has a set of configuration parameters allowing fine tuning of properties like the custom MongoDB collection to which results are saved or the ID of a MapReduce job that must precede the execution of the current job allowing MapReduce jobs to be run in sequence.

Share MapReduce plugins
Once a MapReduce plugin is added to Infinit.e via the Plugin Manager it can be shared across multiple users and MapReduce jobs helping to reduce the amount of time spent developing and managing plugins.

Append to, or age out, old results
The MapReduce job configuration allows developers to specify whether the Reduce function of the plugin overwrites the existing data in the job's custom MongoDB collection or appends to the existing dataset allowing analysis of the data over time.

One of the biggest advantages the Infinit.e Hadoop integration provides to organizations is the ability to integrate with any standard Hadoop distribution including the Hortonworks Data Platform.

HOW INFINIT.E AND HADOOP EXECUTE A MAPREDUCE JOB

The following section is a brief introduction to the mechanics of how Infinit.e and Hadoop work in together to execute a MapReduce job.

Launching the MapReduce job:
There are two ways that Infinit.e can execute a MapReduce job (which involves sending the MapReduce Java JAR and an XML based configuration file to the Hadoop JobClient application):

1. One time API call to execute a job
   The Infinit.e API exposes a simple REST based interface that allows users to execute a MapReduce job (from the browser command line or via the MapReduce Plugin Manager) and pass it to Hadoop's JobClient.
2. Scheduled job
   The Infinit.e Core Server monitors the MapReduce jobs that have been added to the system and passes them to Hadoop's JobClient based on the job's configured execution schedule.

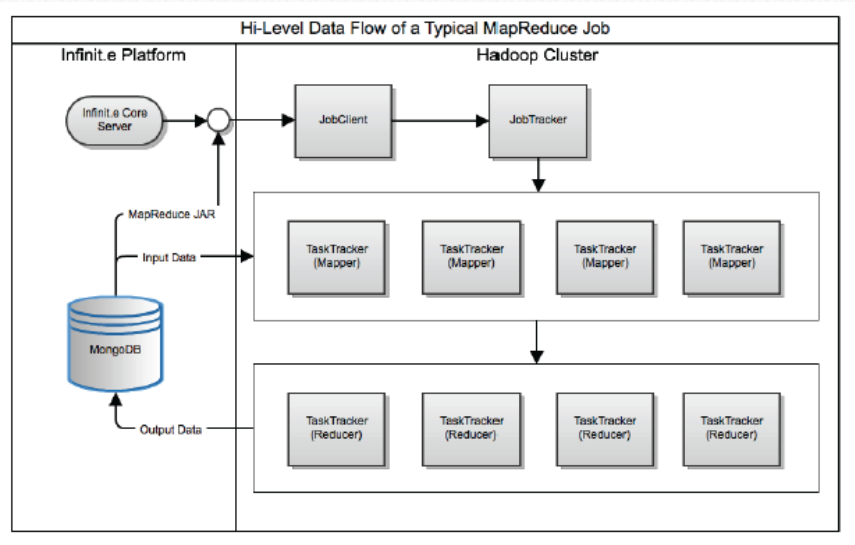
Hadoop JobClient
Hadoop's JobClient is the primary interface that applications can interact with Hadoop to submit MapReduce jobs and track their progress during execution. The Infinit.e Core Server calls JobClient and passes it the Java MapReduce JAR and XML configuration file. JobClient takes the JAR and configuration information, prepares Hadoop for execution (computes InputSplits, copying the JAR and configuration file to Hadoop's file system, etc.) and then submits the job to the JobTracker application.

Hadoop JobTracker
JobTracker is a Hadoop service that manages the process of distributing a MapReduce job across the nodes in a Hadoop cluster including the following steps:

1. Submits work to individual TaskTracker nodes in the cluster;
2. Monitors the TaskTracker nodes assigned to a job and reassigs work if a node fails;
3. Updates the jobs status throughout its life.

Hadoop TaskTracker
The TaskTracker application is a node in the Hadoop cluster that accepts MapReduce tasks (Map, Reduce, Shuffle), executes each task within its own JVM, captures the results of each task, and returns the results or exit code to the JobTracker.

The diagram below illustrates at a high-level the data flow of a MapReduce job executed from the Infinit.e platform.



THE BENEFITS OF INFINIT.E AND HADOOP – AN EXAMPLE

While the aggregations and visualization components that ship with the Infinit.e platform are often all that's required for most document analysis problems, there are more complex data problems that require:

- Access to all raw data, not just pre-aggregated or pre-selected data;
- A high-level programming language like Java with mature, well tested libraries.

For example, the Infinit.e platform, in combination with a Natural Language Processing application like AlchemyAPI makes it trivial to harvest documents from social media sites like Twitter and analyze the harvested entities, associations, and aggregate values. Some NLP tools, like Alchemy API, even give you the ability to extract sentiment around entities from documents allowing analysis of questions related to issues like consumer perception of a new product or voters feelings on political candidate. There are however an even larger set of questions that could be answered if you have the ability to further process and aggregate your data using Hadoop and MapReduce.

Imagine that you have harvested a large amount of Twitter data and used the AlchemyAPI to extract entities, associations, and sentiment data from those Tweets. With this information alone you will be able to answer interesting questions with Infinit.e like:

- Who is Tweeting and how often are they tweeting?
- How are people on Twitter connected to other people (via Tweeting to, about, or re-Tweeting)?
- What are people Tweeting about?
- How do people feel about who or what they are Tweeting about (i.e. sentiment)?

But what if you wanted to answer the more generic question:

Can I get a sense of the general mood of people who Tweet aggregated by geographic location?

Fortunately with Infinit.e and Hadoop MapReduce this is a relatively trivial problem to solve in the following manner:

Create a MapReduce JAR that:

- Has a map method that accepts the results of a MongoDB query and:
- Discards Tweets that are missing location data (latitude and longitude);
- Removes precision from the latitude and longitude values of each tweet so as to make easier to assign documents to a bucket of Tweets in a geographic region;
- Sums the total value of the entity sentiment associated to the tweet
- Returns the document ID, location data, and sentiment value for each Tweet
- Has a reduce method that:
- Aggregates Tweets by geographic location;
- Sums the total sentiment value of the aggregated tweets;
- Returns the results to be written to a custom MongoDB collection.
- Upload the JAR and configure the MapReduce job using Infinit.e's MapReduce Plugin Manager;
- Query the results of your MapReduce job using the Infinit.e API.

The screenshot below illustrates one very simple method that can be used to visualize the results of the MapReduce job described above.

It is important to note that while most data analysis problems requiring MapReduce functionality can be solved in a manner similar to the Twitter Sentiment

example above, the Infinit.e platform allows for more advanced solutions including the chaining of MapReduce jobs (i.e. where the results of one MapReduce job are used as the input for another MapReduce job).

SUMMARY

One of the major challenges facing organizations is simply getting their data into a format that is economically viable to process, query and visualize information with resources at hand. IKANOW's Infinit.e Hadoop integration, as well as other technologies like Natural Language Processing engines helps organizations overcome this barrier by providing an affordable and effective solution for:

- Creating one unified, semi-structured document format from the myriad of structured and unstructured Enterprise data sources making it possible to perform aggregations and processes across the full corpus of documents regardless of their source formatting;
- Open REST based API that makes it easy to integrate with industry standard document analysis and NLP tools;
- And direct integration to Hadoop that greatly minimizes the development time and effort involved in taking advantage of the power of MapReduce.

LEARN MORE ABOUT INFINIT.E

IKANOW is the proud creator of the open source analytics tool Infinit.e. Infinit.e creates agile intelligence through its analytics development environment (ADE) which allows plug and play functionality with additional data sources, visualization widgets and extraction engines.

Get started with Infinit.e by **downloading the open source documentation** or scheduling a **free demonstration of how Infinit.e can help your business or organization** create actionable intelligence from big data.

Feel free to share this eBook and download additional resources at **http://ikanow.com**.

Thank you for downloading our eBook and let us know your thoughts by **contacting us via email or Twitter**.