DSBDA Case study

Name of topic: Case study on Datameer Analytics Solution

Batch name: L9

Group ID: 6

Name of Students:

1. Ajinkya Kulkarni (33126)
2. Yash Kulkarni (33129)
3. Shubham Loya (33133)
4. Sahil Naphade (33140)

Introduction to the topic:

As per Datameer.com, they define their product as “Datameer X: Data Prep for Machine Learning.”. They are providing a solution to shorten the machine learning workflows, by using Hadoop. The company had a humble beginning of spending several years implementing custom Hadoop solutions even before Yahoo! even came into picture and Hadoop was called as Nutch. And finally, they developed a self-service big data analytics tool.

Therefore, Datameer is a self-service analytics platform that can integrate data from any source, size or variety. The platform is used to empower employees rather than relying solely on IT for big data insights in big companies. Users can develop on their laptop, test with their work group, deploy to their company, and scale with their needs: on their laptop, cluster, or server.

Thus, in this study we aim to study the working of Datameer Analytics solution, its use cases, integration with Hadoop, AWS and Azure platform, installation steps.

Overview:

Datameer analytics solution is a self-service big data analytics tool, developed by Datameer, headquartered in California, US, established in 2009.

The product allows for Data Integration, Data visualization, Dynamic data management and self-service analytics on an open infrastructure, Hadoop. It is a SaaS type of cloud application, aimed to make everyone from small business to largest MNC’s to make data-driven decisions regardless of their level of technical expertise of higher management. Datameer is built natively on Hadoop and leverages inexpensive hardware so that users can integrate and analyze all of their data in its raw form without the need for IT to perform ETL and pre-model the data.

As of now, the latest version of Datameer is ‘Datameer X’, whereas Datameer 1.0 came out in 2010.

Purpose:

Datameer allows all its consumers, ranging from smallest businesses to largest MNCs to facilitate the data-driven decision making regardless of their technical proficiency. They allow their customers to make smarter decisions, by delivering feature-rich machine learning preparation functions and large-scale data exploration options to data scientists by a mechanism as simple as point-and-click.

Datameer has long been a leader in two areas of shaping data for machine learning and large-scale data exploration with its unique Smart Analytics™ and Visual Explorer features. Datameer X extends this leadership with:

- New point-and-click machine learning encoding functions—including OneHot, Ordinal, and Date and Number Binning—that automate the process of shaping data to feed machine learning models
- Pivot table capabilities (similar to Excel) that allow users to explore and organize data into tables at scale with support for billions of records, hundreds of attributes, and thousands of categorical values
- A new production execution mode which allows operationalized jobs to run up to 30 percent faster
- Enhanced integration with external data sources and destinations such as with new connectors for Google BigQuery and Tableau Hyper.

Datameer works on all popular open source and commercial Hadoop distributions including Apache Hadoop, Cloudera, EMC, Hortonworks, IBM and MapR. Datameer is the only big data analytics application built natively on Hadoop. Datameer combines the traditional three processes and three separate vendors into one easy-to-use application across data integration, analysis, and visualization. Datameer is an open and extensible solution. Users can: - choose any Hadoop distribution (Cloudera, Hortonworks, MapR, IBM, etc.) - choose any data sources (Twitter, CRM databases, web logs, DBMS, POS info, etc.) - work with over 200+ pre-built analytic functions - Use the APIs to build custom data connectors and custom analytic functions.

Need:

Datameer is used in all the cases, inclusive but not limited to Fraud detection, Operational analytics, Customer analytics, connected home and many more. Also, Datameer provides a robust set of data management capabilities that include data retention policies, automated partitioning, compaction and compression, and incremental data loading. On ingestion, Datameer can filter out corrupt, dirty and incomplete data, aiding your data cleansing process.

It is also useful in industry wide application, including Asset and Material usage, Billing analytics, clinical trials, Financial Transactional analysis.

Thus, it is an all-comprehensive solution intended to ease the working of any industry, in any use case. It uses power of Hadoop, to effectively manage big amount of structured as well as unstructured data to visualize, analyze and predict the data.

It helps in Customer behavior analytics powered by big data helping organizations optimize that customer journey. It helps in fraud mitigation by Big data analytics identifying patterns deep in the data to identify fraud and aggregate large volumes of information to make regulatory reporting much faster. Use of larger data sets makes identification of fraud patterns and algorithms more accurate.

Services provided:

Datameer provides ease of integration of big, unstructured data, to aid in analysis, visualization and conclusion drawing easier.

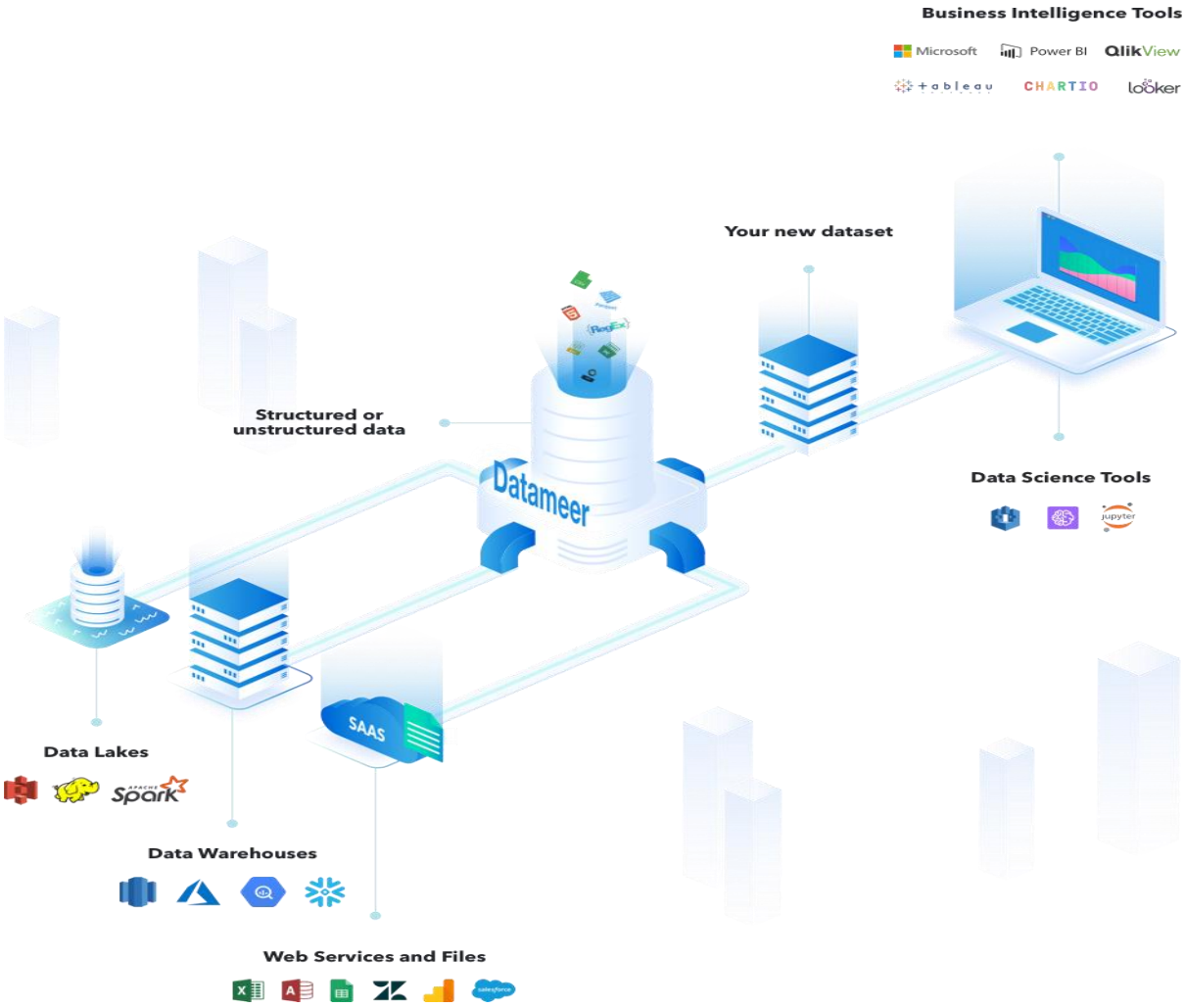

Fig 1. High-level overview of 'Datameer X' [1]

As we can see in the infographic above, Datameer takes the information from all the sources, like Files, Warehouses and data lake, and integrates it to provide a new dataset, which can then be used for data science tools for further processing, thus living up to its definition as a tool for Machine Learning preparation. Thus, Datameer allows for all the features required to manage a data-driven business, like integration from various sources, preparing it for direct usage in data analysis tools. It allows for

[1] Datameer

storage of data using Hadoop, cleaning, integrating, visualization and conclusion of the details.

It provides following features:

- Data integration
- Data visualization
- Dynamic data management
- Open infrastructure
- Pre-built applications
- Self-service analytics

Rightfully, it is the only full-scale solution build only on Hadoop platform, and it excels at it.

Installation steps[2]:

1. Create the Datameer User

```
./usr/sbin/groupadd --system datameer
./usr/sbin/useradd --system --create-home --gid datameer datameer
```

2. Create Directories for Application, Cache, Logs, and Temporary Files

```
mkdir -p /opt/datameer
chown -R datameer:datameer /opt/datameer
mkdir -p /var/cache/datameer
chown -R datameer:datameer /var/cache/datameer
mkdir -p /var/log/datameer
chown -R datameer:datameer /var/log/datameer
mkdir -p /tmp/datameer
chown -R datameer:datameer /tmp/datameer
```

3. Switch the User and Change the Working Directory

```
su - datameer
cd /opt/datameer
```

[2] Installation guide to Datameer

4. Download and unzip Datameer

```
curl -s -k -o Datameer-<package>.zip "https://download.datameer.com.s3.amazonaws.com/releases/Datameer-<version>/<dist>/Datameer-<package>.zip?<AWSproperties>" ; unzip Datameer*
```

5. Download and Install the MySQL Database JDBC Connector

By default, the Datameer application runs with an HSQL file database that is created on the local filesystem under das-data/database/hsql-db. If you are setting up Datameer for production use, Datameer strongly recommends using MySQL instead of the HSQL file database.

Installing JDBC:

```
# Lookup latest JDBC driver version
JDBCDRV="$(curl -s -k 'https://dev.mysql.com/downloads/connector/j/' | grep -o -m 1 'mysql-connector-java.*zip')"
# Download latest JDBC driver version
curl -s -L0 -k -O "https://dev.mysql.com/get/Downloads/Connector-J/${JDBCDRV}"
# Unzip driver package
unzip mysql-connector* -d etc/custom-jars
# Move only the necessary JAR file
mv etc/custom-jars/mysql-connector*/*bin.jar etc/custom-jars
# Clean up
rm -rf etc/custom-jars/mysql-connector-java-?.?.??

Checking installation
echo $JDBCDRV
ll etc/custom-jars
```

6. Configure Datameer for MySQL Database

Datameer service depends on the MySQL database. The MySQL database is used for writing to workbooks, permission changes, job execution, scheduling, and more. To function properly, a response time should be between ten and twenty milliseconds. To run the application in MySQL mode, the following changes need to be implemented. **As of Datameer 7.4:** MariaDB is supported as an alternative to MySQL.
Check database connection:

Connection check

```
mysqladmin version
mysqladmin ping
mysqladmin status
echo q | telnet -e q `hostname` 3306
nc -z -w1 `hostname` 3306
```

You can follow up later with using the Check if the Datameer Application Database is Running and Accessible article.
Initialize application database:

Initialize database

```
mysql -uroot -p < bin/mysql-init.sql
mysql -uroot -p dap < bin/create-tables.sql
```

Configure for an enterprise environment:

etc/das-env.sh

```
# Create a backup of the original configuration file
cp etc/das-env.sh etc/das-env.sh.original
# Change the deploy mode
sed -i "s/\(DAS_DEPLOY_MODE=\).*\$/\1live/" etc/das-env.sh
# Uncomment the database name you will be using
sed -i '/#.*DATAMEER_DB_NAME=/s/^#//' etc/das-env.sh
# Uncomment the user that the application should be started at
sed -i '/#.*DAS_USER=/s/^#//' etc/das-env.sh
# Specify the maximum size of the memory allocation pool
sed -i 's/Xmx2048m/Xmx4096m/' etc/das-env.sh
# Create a log of changes made
diff -e etc/das-env.sh.original etc/das-env.sh > changes.das-env.sh
```

7. Installing the License

If you have already received a license from Datameer, copy the license file to: `$INSTALL_LOCATION/das-<version>/etc/license` or log into Datameer from the user interface and go to the **Admin** tab. Select **License** from the side menu and upload your license here.

The file must be named *license.lic* or it won't work.

If you don't have a license, email license@datameer.com.

If you attempt to launch the application without a license, you receive an message asking you to contact Datameer for a license file. Include the MAC address listed in this error message when contacting Datameer for a license. License information and usage can be viewed at any time by clicking the **Admin** tab and selecting **License**.

See License Information for information on how to update the license and for details about volume-based licensing.

At this time Datameer has been installed.

8. Start Datameer

Start Datameer

```
# Start the Datameer service
./bin/conductor.sh start
# Check the process ID (PID)
ps -ef | grep -i "java.*jetty.*datameer" | grep -v grep | tr -s " " | cut -
d " " -f2
# Monitor the process booting and the log files
cat logs/jvm-stdout.log; sleep 3; tail -F logs/`date +"%Y_%m_%d"`.st
```

9. Stop Datameer

Stop the Datameer service.
Working within the `current` installation directory, use the following commands:

Start Datameer

```
# Stop the Datameer service
./bin/conductor.sh stop
# Monitor the process shutting down
cat logs/jvm-stdout.log; sleep 3; tail -F logs/`date +"%Y_%m_%d"`.
```

Costing:

Datameer is not a free-for-use product. It is an annual subscription based product, with 3 options:

1. Personal ($300pa)
2. Workgroup edition ($19,188pa)
3. Enterprise edition (Contact vendor)

Flexibility:

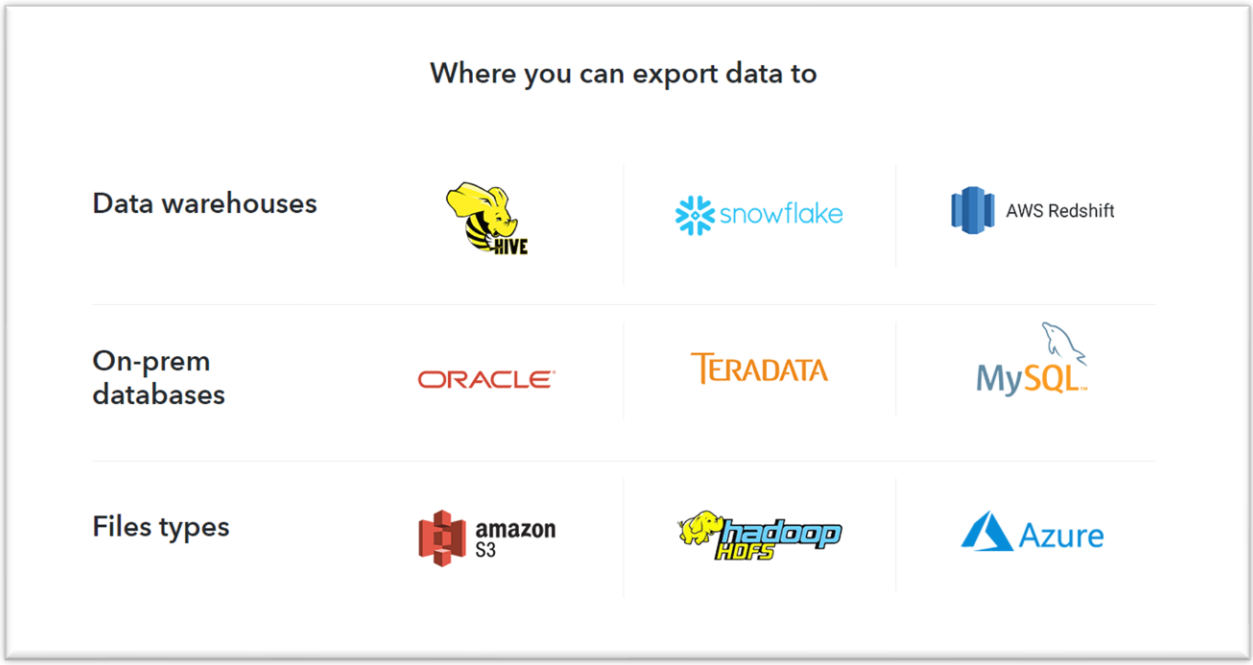

Fig 2. The data in Datameer can be exported to

How Datameer works:

Datameer acts as a job compiler or code generator like Hive. This means every function, filter or join that the user designs in the spreadsheet will be translated into native Tez code. Tez is great for splitting up workloads into smaller pieces. To do so, Datameer compiles a job for a Hadoop cluster, where it is sent to be executed. After the job is compiled and sent to the cluster Datameer does not control job execution, and can only receive the telemetry metrics provided by the cluster's services. The job will run with any scheduling settings and use resources granted by the scheduler.

All users working with Datameer's Excel-like User Interface (UI) are generating a Java program for distributed computing on the cluster backend. This high level of abstraction is one of the key features that makes Datameer such an outstanding technology. However, this approach does mean that business users need to keep in mind the types of problems every programmer deal with, i.e. data types, memory, and disk usage.

This separates analytics work into two stages. First, the design/edit time and second the execution/runtime of a data link/import job/workbook. Both stages are located on different parts within your distributed computing system (cluster).

Use and customers:

Datameer is primarily used in Machine learning data prep. It is an official partner of Amazon AWS.

It has a very wide range of customers, all depending on data-driven decisions[3]:

a. Siemens
b. Deutsche Bank
c. Scotiabank
d. HCSC
e. Anthem BlueCross
f. Citi Bank
g. National instruments

And many more….

Advantages:

1. Built natively on Hadoop, so extremely scalable.
2. Easy and intuitive to use by Data scientists.
3. Allows fast and efficient integration of structured as well as unstructured data.
4. Effective and powerful machine learning prep of data.
5. Seamless integration with commercial cloud platforms like Azure, AWS and even on-premise cloud.
6. Datameer Enterprise Edition is priced on a data plan model of how much new data is brought into Datameer per year, so it is reasonable.

Conclusion:

Thus, this is an all-inclusive of study of Datameer, the only SaaS big data analytics tool built natively on Hadoop. It has a very high integrability with all the commercial cloud platforms. Though it is not cheap to use, it is an excellent choice for small businesses to MNCs to handle their data, owing to high level machine learning data prep capability, and a very high up-time.

[3] Source : https://www.datameer.com/customers/