Welcome to “Cloud Based Tools for Data Science.”

After watching this video, you will be able to:

Describe how commercial cloud tools support data science tasks, and

Explain how integration provides the ability to use the same tools for multiple tasks.

Let’s again look at the overview of different tool categories.

Since cloud products are a newer species, they follow the trend of having multiple tasks

integrated in tools.

This integration is applicable for the tasks marked green in the diagram.

Let’s start with the fully integrated visual tools category.

Since these tools introduce a component where large-scale execution of data science workflows

happens in compute clusters, we have changed the title and added the word “Platform.”

These clusters are composed of multiple server

machines, transparently for the user, in the background.

Watson Studio and Watson OpenScale cover the complete development life cycle for all data

science, machine learning, and artificial intelligence (AI) tasks.

Another example is Microsoft Azure Machine Learning.

It is also a full cloud-hosted offering supporting the complete development life cycle of all

data science, machine learning, and AI tasks.

And finally, another example is H2O Driverless AI.

Although it is a product you download and install, there exists a one-click deployment

for the standard cloud service providers.

Since the cloud provider does not do operations and maintenance, as with Watson Studio, Open

Scale, and Azure Machine Learning, this delivery model should be distinct from Platform or

Software as a service - PaaS or SaaS.

In data management, with some exceptions, software-as-a-service (SaaS) versions of existing

open source and commercial tools exist.

The cloud provider operates the tool for you in the cloud.

For example, the cloud provider operates the product by backing up your data and configuring

and installing updates.

Some proprietary tools are only available from a single cloud provider.

One example of such a service is Amazon Web Services DynamoDB, which is a NoSQL database.

It allows storage and retrieving data in a key-value or a document store format.

The most prominent document data structure is JSON.

Another flavor of such a service is Cloudant, which is a database as a service offering.

But, in the background, it is based on the open-source Apache CouchDB.

The advantage is that complex operational tasks like updating, backup, restoring, and

scaling are done by the cloud provider.

However, the Cloudant service offering is compatible with CouchDB.

Therefore, the application migrates to another CouchDB server without making any changes

to the application.

IBM offers Db2 as a service as well.

It is an example of a commercial database made available as a SaaS offering in the cloud,

taking away operational tasks from the user.

Now let’s discuss commercial data integration tools that include extract, transform, and

load (ETL) tools and extract, load, and transform (ELT) tools.

It means the transformation steps are not done by a data integration team but are pushed

toward the domain of the data scientist or data engineer.

Two commercial data integration tools widely used are Informatica Cloud Data Integration

and IBM’s Data Refinery.

Data Refinery is part of IBM Watson Studio.

It allows transforming large amounts of raw data into consumable, quality information

in a spreadsheet-like user interface.

So, the market for cloud data visualization tools is huge, and every major cloud vendor has one.

An example of a smaller company offering a cloud-based data visualization tool is Datameer.

IBM offers its famous Cognos Business intelligence suite as a cloud solution.

IBM Data Refinery also offers data exploration and visualization functionality in Watson Studio.

Again, those are examples of a rapidly changing and growing commercial ecosystem among many

established, and emerging vendors.

In Watson Studio, various visualizations depict data for better understanding.

An example is this 3D bar chart that visualizes a target value on the vertical dimension that

is dependent on two other values in the horizontal dimensions.

You can use colors to visualize the third dimension.

Another data visualization is hierarchical edge bundling that depicts correlations and

affiliations between entities.

If sufficient, a classic bar chart can do the job as well.

A 2D scatter plot with a heat map shows two dependent data fields on the y-axis with different

color intensities.

A tree map shows the distribution of subsets within a set.

The famous pie chart does the same but in a non-hierarchical manner.

Finally, a word cloud pops out significant terms in a document corpus.

Model building can be done using a service.

One example of a service is Watson Machine Learning.

Watson Machine Learning can train and build models using various open-source libraries.

Google has a similar service on their cloud called AI Platform Training.

Every cloud provider has a solution for this task.

Model deployment in commercial software is usually tightly

integrated into the model-building process.

Here is an example of the SPSS Collaboration and Deployment Services which can be used to deploy any asset created

by the SPSS software tools suite.

The same holds for other vendors.

In addition, commercial software can export models in an open format.

For example, SPSS Modeler supports exporting models as Predictive Model Markup Language

(PMML), which other commercial and open software packages can read.

In addition, Watson Machine Learning deploys

a model and makes it available to consumers using a REST interface.

Amazon SageMaker Model Monitor is an example of a cloud tool to monitor deployed machine

learning and deep learning models continuously.

Every major cloud provider has similar tooling.

Another tool for model monitoring is Watson OpenScale.

Everything marked in green can be done using Watson Studio and OpenScale.

Play video starting at :7:6 and follow transcript7:06

In this video, you’ve learned:

Watson Studio and Watson OpenScale,

cover the complete development life cycle for all data science, machine learning, and AI tasks.

In data management, with some exceptions, there exists a software-as-a-service (SaaS) version of

existing open-source and commercial tools.

Two commercial data integration tools widely used are Informatica Cloud Data Integration and IBM’s Data Refinery.

An example of a cloud-based data visualization tool is Datameer and IBM’s

Cognos Business intelligence suite.

Model building can be done using a service such as Watson Machine Learning.

Amazon SageMaker Model Monitor is an example of a cloud tool to monitor deployed machine-learning and deep learning models continuously.

Previously saved note: For example, the cloud provider operates the product by backing up your data and configuring and installing updates.. Press [CTRL + D] to delete the note

Transcript language: English

​

Welcome to “Cloud Based Tools for Data Science.”

After watching this video, you will be able to:

Describe how commercial cloud tools support data science tasks, and

Explain how integration provides the ability to use the same tools for multiple tasks.

Let’s again look at the overview of different tool categories.

Since cloud products are a newer species, they follow the trend of having multiple tasks

integrated in tools.

This integration is applicable for the tasks marked green in the diagram.

Let’s start with the fully integrated visual tools category.

Since these tools introduce a component where large-scale execution of data science workflows

happens in compute clusters, we have changed the title and added the word “Platform.”

These clusters are composed of multiple server

machines, transparently for the user, in the background.

Watson Studio and Watson OpenScale cover the complete development life cycle for all data

science, machine learning, and artificial intelligence (AI) tasks.

Another example is Microsoft Azure Machine Learning.

It is also a full cloud-hosted offering supporting the complete development life cycle of all

data science, machine learning, and AI tasks.

And finally, another example is H2O Driverless AI.

Although it is a product you download and install, there exists a one-click deployment

for the standard cloud service providers.

Since the cloud provider does not do operations and maintenance, as with Watson Studio, Open

Scale, and Azure Machine Learning, this delivery model should be distinct from Platform or

Software as a service - PaaS or SaaS.

In data management, with some exceptions, software-as-a-service (SaaS) versions of existing

open source and commercial tools exist.

The cloud provider operates the tool for you in the cloud.

For example, the cloud provider operates the product by backing up your data and configuring

and installing updates.

Some proprietary tools are only available from a single cloud provider.

One example of such a service is Amazon Web Services DynamoDB, which is a NoSQL database.

It allows storage and retrieving data in a key-value or a document store format.

The most prominent document data structure is JSON.

Another flavor of such a service is Cloudant, which is a database as a service offering.

But, in the background, it is based on the open-source Apache CouchDB.

The advantage is that complex operational tasks like updating, backup, restoring, and

scaling are done by the cloud provider.

However, the Cloudant service offering is compatible with CouchDB.

Therefore, the application migrates to another CouchDB server without making any changes

to the application.

IBM offers Db2 as a service as well.

It is an example of a commercial database made available as a SaaS offering in the cloud,

taking away operational tasks from the user.

Now let’s discuss commercial data integration tools that include extract, transform, and

load (ETL) tools and extract, load, and transform (ELT) tools.

It means the transformation steps are not done by a data integration team but are pushed

toward the domain of the data scientist or data engineer.

Two commercial data integration tools widely used are Informatica Cloud Data Integration

and IBM’s Data Refinery.

Data Refinery is part of IBM Watson Studio.

It allows transforming large amounts of raw data into consumable, quality information

in a spreadsheet-like user interface.

So, the market for cloud data visualization tools is huge, and every major cloud vendor has one.

An example of a smaller company offering a cloud-based data visualization tool is Datameer.

IBM offers its famous Cognos Business intelligence suite as a cloud solution.

IBM Data Refinery also offers data exploration and visualization functionality in Watson Studio.

Again, those are examples of a rapidly changing and growing commercial ecosystem among many

established, and emerging vendors.

In Watson Studio, various visualizations depict data for better understanding.

An example is this 3D bar chart that visualizes a target value on the vertical dimension that

is dependent on two other values in the horizontal dimensions.

You can use colors to visualize the third dimension.

Another data visualization is hierarchical edge bundling that depicts correlations and

affiliations between entities.

If sufficient, a classic bar chart can do the job as well.

A 2D scatter plot with a heat map shows two dependent data fields on the y-axis with different

color intensities.

A tree map shows the distribution of subsets within a set.

The famous pie chart does the same but in a non-hierarchical manner.

Finally, a word cloud pops out significant terms in a document corpus.

Model building can be done using a service.

One example of a service is Watson Machine Learning.

Watson Machine Learning can train and build models using various open-source libraries.

Google has a similar service on their cloud called AI Platform Training.

Every cloud provider has a solution for this task.

Model deployment in commercial software is usually tightly

integrated into the model-building process.

Here is an example of the SPSS Collaboration and Deployment Services which can be used to deploy any asset created

by the SPSS software tools suite.

The same holds for other vendors.

In addition, commercial software can export models in an open format.

For example, SPSS Modeler supports exporting models as Predictive Model Markup Language

(PMML), which other commercial and open software packages can read.

In addition, Watson Machine Learning deploys

a model and makes it available to consumers using a REST interface.

Amazon SageMaker Model Monitor is an example of a cloud tool to monitor deployed machine

learning and deep learning models continuously.

Every major cloud provider has similar tooling.

Another tool for model monitoring is Watson OpenScale.

Everything marked in green can be done using Watson Studio and OpenScale.

Play video starting at :7:6 and follow transcript7:06

In this video, you’ve learned:

Watson Studio and Watson OpenScale,

cover the complete development life cycle for all data science, machine learning, and AI tasks.

In data management, with some exceptions, there exists a software-as-a-service (SaaS) version of

existing open-source and commercial tools.

Two commercial data integration tools widely used are Informatica Cloud Data Integration and IBM’s Data Refinery.

An example of a cloud-based data visualization tool is Datameer and IBM’s

Cognos Business intelligence suite.

Model building can be done using a service such as Watson Machine Learning.

Amazon SageMaker Model Monitor is an example of a cloud tool to monitor deployed machine-learning and deep learning models continuously.

Previously saved note: For example, the cloud provider operates the product by backing up your data and configuring and installing updates.. Press [CTRL + D] to delete the note

Transcript language: English

​