Google Cloud Computing Foundation Course
Evan Jones
Technical Curriculum Developer
Google Cloud

Lecture-71
Leverage Big Data Operations with Cloud Dataproc

In this topic learn about how cloud data proc provides a fast, easy, cost-effective way to run Apache Hadoop in Apache spark which are open source Big Data technologies that support big data operations.

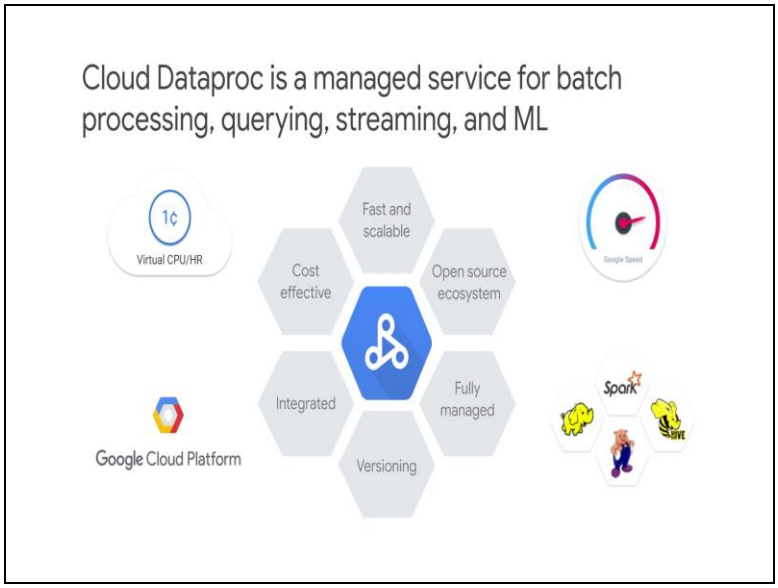
(Refer Slide Time: 00:14)

Hadoop and Spark are open source technologies that often form the backbone of big data processing. Hadoop is a set of tools and technologies which enable a cluster of computers to store and process large volumes of data it intelligently ties together individual computers in a cluster to distribute the storage in the processing of that data. Apache spark is a unified analytics engine for large-scale data processing and achieves both high performance for both batch and streaming data.

Cloud data proc is a managed Spark and Hadoop service that lets you take advantage of the open source data tools for batch processing, querying, streaming and machine learning. Cloud Data proc automation helps you quickly create those clusters manage them easily and because clusters are typically run ephemerally meaning short-lived you will save money as they are turned off

when you do not need that processing power anymore. Let us take a look at the key features of cloud dataproc.

(Refer Slide Time: 01:17)



Cloud dataproc is priced at 1 cent per virtual CPU per cluster per hour on top of any other GCP resources that you use. In addition cloud dataproc clusters can include preemptable instances that have lower compute prices. You use and pay for things only when you need them and not when you do not. Cloud Dataproc clusters are quick to start to scale and to shutdown which each of these operations taking 90 seconds or less on average.

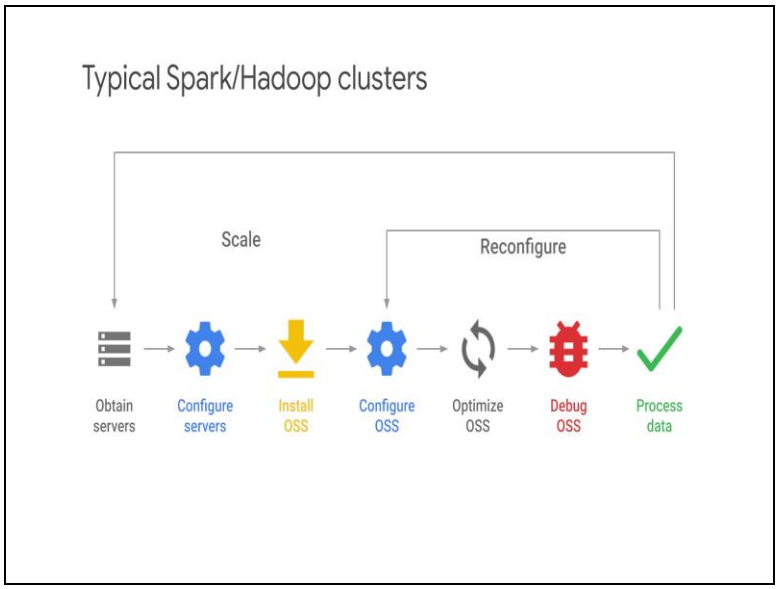
Clusters can be created and scaled quickly with a variety of virtual machine sizes, types, number of nodes and networking options. You can use spark and Hadoop tools libraries and documentation with cloud dataproc. Cloud dataproc provides frequent updates to native versions of spark, Hadoop, pig and hive so there is no need to learn tools or the API's it is possible to move your existing projects or your ETL pipelines it is a Google cloud without redevelopment.

You can easily interact with clusters and spark or Hadoop jobs without the assistance of an administrator or special software for the GCP console the cloud SDK or the cloud dataproc REST API. When you are done with the cluster simply turn it off so money is not spent on an idle cluster. Image versioning allows you to switch between different versions of Apache spark, Apache Hadoop and other tools.

The built-in integration with cloud storage BigQuery and cloud BigTable ensures data will never be lost even when your cluster turns down. This together with stackdriver logging and stockdriver monitoring provides a complete data platform and not just a spark or a Hadoop cluster. For example you can use cloud dataproc to effortlessly ETL terabytes of raw log data directly into bigquery for your business reporting needs.

So how does cloud dataproc work spin up a cluster when needed for example to answer a specific query or run a specific ETL job.

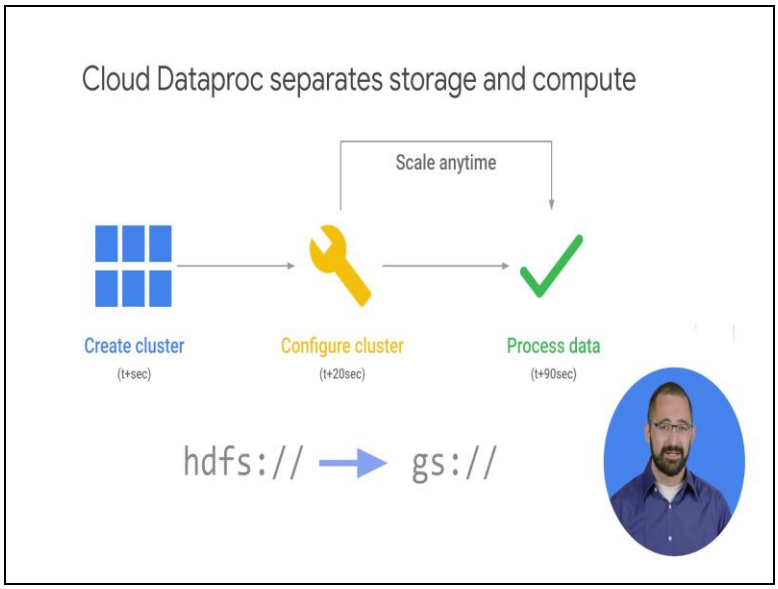
(Refer Slide Time: 03:25)



The architecture depicted here provides insight into how the cluster remains separate yet easily integrates with other important functionalities for example logging via stackdriver and cloud big table instead of Hbase. This contributes to the ability for Cloud Dataproc to run ephemerally and therefore efficiently and cost-effectively. The Cloud Dataproc approach allows users to use Hadoop, spark, hive and pig when they need it.

Again as we mentioned it only takes 90 seconds on average from the moment users request the resources before they can submit their first job. What makes this possible is this separation of storage and compute which is a real game changer.

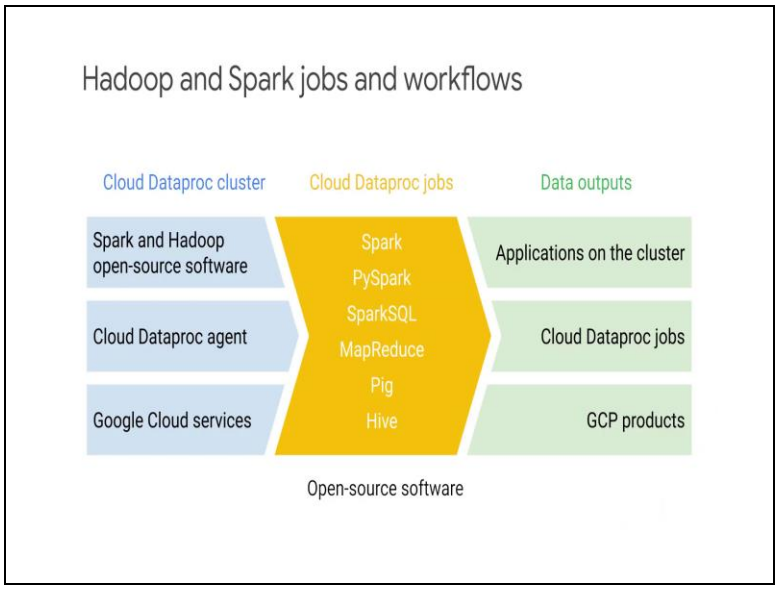
(Refer Slide Time: 04:08)



With their traditional approach typical, on-premise clusters their storage and their hard drives are attached to each of the nodes in the cluster. If the cluster is not available to a maintenance neither is the storage. Since the storage is attached to the same computing nodes as those that do the processing there is often a contention for resources for example input and output bottlenecks on the cluster. Cloud dataproc on the other hand relies on storage resources being separated from those compute resources.

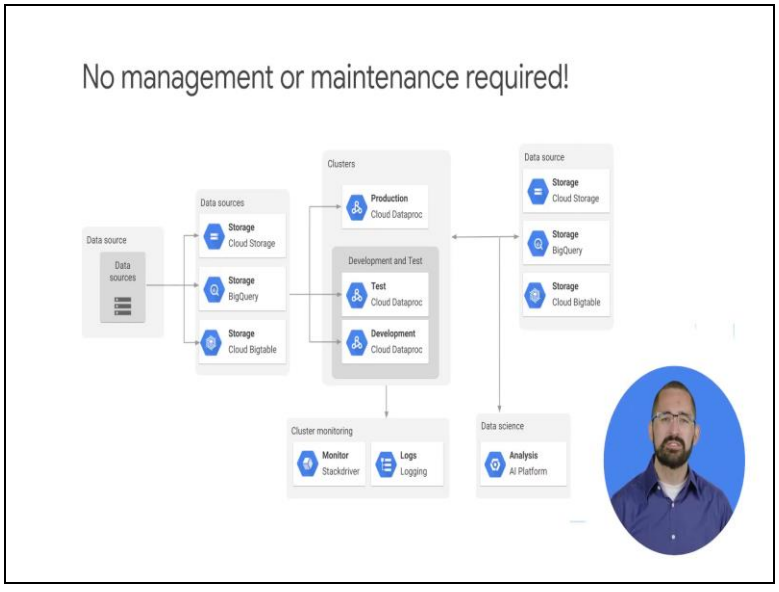
Files are stored on Google Cloud Storage or on the Google Cloud Storage connector meaning that using Google Cloud Storage instead of HDFS is as easy as changing the prefix in the scripts from HDFS to GS or Google storage.

(Refer Slide Time: 04:53)



Also consider cloud dataproc in terms of Hadoop and spark jobs and workflows. The workflow template allows users to configure and execute one or more jobs. It is important to remember that beyond making the process easier for example by allowing the user to focus on jobs and view the logs and stackdriver. They can always access the Hadoop components and applications such as the Yarn Web UI running on their cloud dataproc cluster if they wanted to.

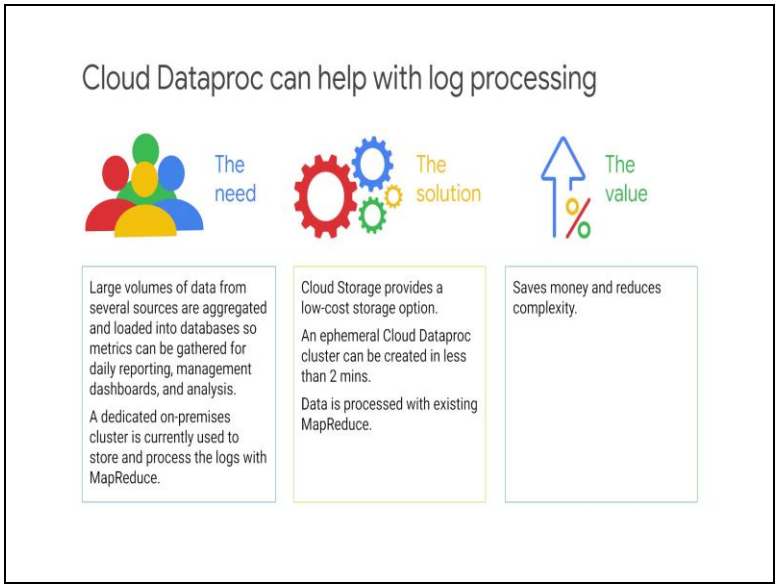
(Refer Slide Time: 05:23)



To run a cluster when needed for a given job to answer a specific query this architecture shows what is possible and how it can integrate with managed services running outside the cluster. For

example logging and monitoring through stackdriver or cloud BigTable instead of traditional HBase.

(Refer Slide Time: 05:40)

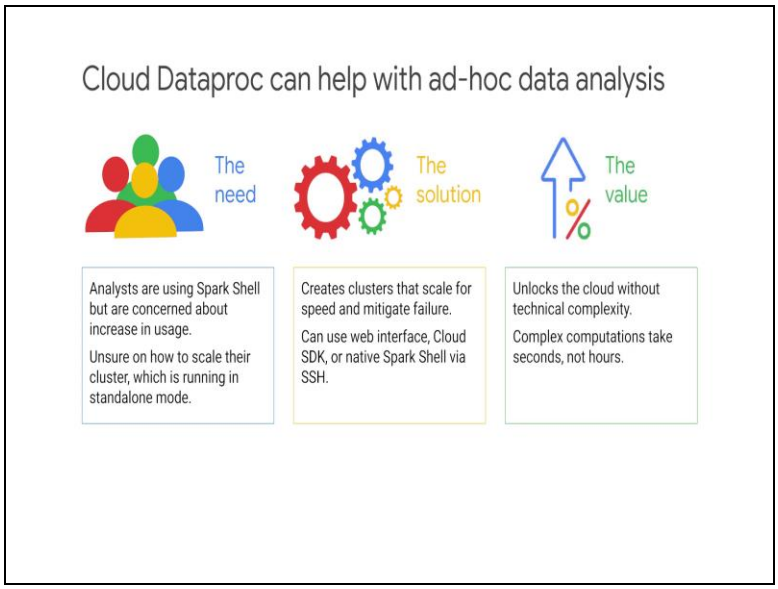


Let us look at a few of those use cases. Starting with how cloud data proc can help with log processing. In this example a customer processes 50 gigabytes of text log data per day from several sources to produce aggregated data that is then loaded into databases from which metrics are then gathered for things like daily reporting, management dashboards and analysis. |Up until now they have used a dedicated on-premises cluster to store and process the logs with MapReduce.

So what is the solution? Firstly cloud storage can act as a landing zone for the log data at low cost. A cloud dataproc cluster can then be created in less than two minutes to process this data with their existing MapReduce. Once completed the cloud dataproc cluster can be removed immediately it is not needed anymore. In terms of value instead of running all the time and incurring costs when it is not used.

Cloud dataproc only runs to process those logs which saves money and reduces your overall complexity.

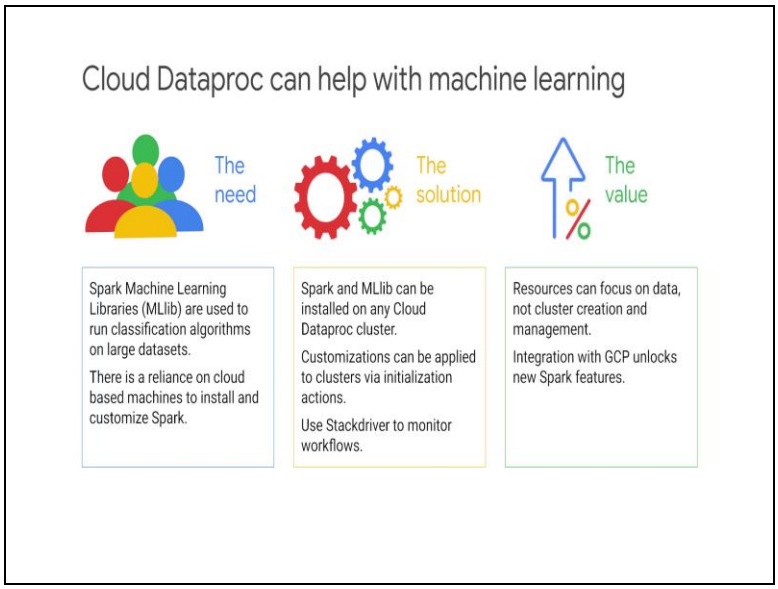
(Refer Slide Time: 06:47)



The second use case looks at how cloud data proc can help with adhoc analysis. In this organization and let us rely and are comfortable with using spark shell however their IT department is concerned about the increase in usage and how to scale their cluster which is running in standalone mode. As a solution cloud dataproc can create clusters that scale for speed and mitigate any single point of failure.

Since the cloud dataproc supports spark, spark sequel and Pi spark they could use the web interface cloud SDK or the native spark shell via SSH. In terms of value cloud data product quickly unlocks the power of the cloud for anyone without adding technical complexity. Running complex computations now take seconds instead of minutes or hours on premise.

(Refer Slide Time: 07:41)



The use case in this third example looks at how cloud dataproc and help with machine learning. In this example a customer uses spark machine learning libraries to run classification algorithms on very, very large data sets. They rely on cloud based machines will they install and customize spark because spark and the machine learning libraries can be installed on any cloud dataproc cluster. The customer can save time by quickly creating cloud dataproc clusters.

Any additional customization can be applied easily to the entire cluster through what are called initialization actions. To keep an eye on workflows they can be used with the built-in cloud logging and monitoring solutions. In terms of value resources can be focused on the data with cloud dataproc not spent on things like cluster creation and management. Also integrations with other new GCP products can unlock new features for your spark clusters.