IBM Analytics
Solution Brief

IBM Analytics Engine

Empower your data scientists with an agile cloud environment for Hadoop and Spark analytics

Highlights

- Provision Apache Hadoop and Apache Spark clusters from a single point of control
- Spin up and shut down clusters within minutes, whenever you need them
- Separate compute and storage infrastructure to improve maintainability and reduce costs
- Integrate IBM® Watson® Data Platform tools to create a complete ecosystem for data science

Delivering an agile architecture for data science

IBM Analytics Engine is a cloud-based service that enables data scientists to rapidly provision, manage, run and retire Apache Hadoop and Apache Spark clusters. The solution is designed to solve the key pain points that organizations currently experience as they try to build up their big data analytics capabilities.

Identifying data science pain points

In many cases, businesses that have invested in Hadoop-based data infrastructure are struggling to achieve their goals, for a number of reasons.

First, implementing and maintaining a Hadoop or Spark cluster is a complex task, requiring a deep understanding of both the underlying infrastructure and the individual software components that make up the Hadoop ecosystem. Data scientists often lack this type of expertise—and even if they have it, cluster management is a distraction from their core role of exploring new data sets and building predictive models.

Second, there is a lack of enterprise-class tools for managing multiple Hadoop and Spark clusters across an organization, or embedding them seamlessly into the broader data science workflow. As a result, individual data science teams often select whatever tools work best for their immediate use case. This approach leads to fragmentation, and makes it difficult for teams to collaborate with each other or with the rest of the business.

Third, most data lakes have been designed to use Hadoop clusters as a permanent store for big data because it is faster to analyze data where it resides than to move it to a new environment for analysis. In practice, however, the problems of using a Hadoop cluster as a persistent data store tend to outweigh the benefits. These problems include:

- **Under-utilized hardware.** Each node in a cluster requires powerful processors, but they are only used when a data scientist needs to access the node's data. The rest of the time, the processors lie idle.
- **Complex maintenance.** Whenever you need to update a Hadoop component, you must take the entire cluster offline to perform the upgrade.
- **Reliability concerns**. Because the cluster is responsible both for data storage and computation, any failure has potential for data loss or corruption.
- **Reduced flexibility.** The number of clusters and the types of nodes must be defined in advance. If the data science workload changes over time, the data lake cannot easily be adapted to meet demand without paying a high price.

Introducing IBM Analytics Engine

Scalability

IBM Analytics Engine takes a different approach from traditional data lakes. Instead of using Hadoop as both a computation engine and a persistence layer for long-term data storage, it splits compute and storage into two separate concerns.

Rather than storing data on the local disks of each node in a cluster, IBM Analytics Engine takes advantage of cloud object storage technology. The data lives permanently in the object store, and replicating it to multiple data centers to protect against loss or corruption is simply a software option for the user. Moreover, the cost of storing data on purpose-built object storage is lower than storing it on a Hadoop cluster with high compute or memory requirements.

The separation of storage and compute resources enables both to scale independently with the needs of the organization. As data volumes grow, you can add more storage; and as the number and scope of data science projects increase, you can add more nodes to your clusters, or spin up clusters with different node-sizes.

Flexibility

By delegating responsibility for permanent storage to an object store, IBM Analytics Engine also enables a much more flexible approach to cluster management. Instead of keeping a fixed-size cluster online at all times, users can provision and deprovision appropriately sized clusters whenever they need them.

For example, when a data scientist wants to start a new project, they can simply log into IBM Analytics Engine, select the number of nodes they need, choose the Hadoop pack or Spark pack, and link the cluster to the object store that contains the data sets they want to analyze.

Within a few minutes, the cluster will be provisioned and ready for the data scientist to start running their Spark or Hadoop jobs. And once those jobs are complete, the user can simply save the results to the object store, and then spin down the cluster.

This is possible because instead of running on bare-metal servers, the IBM Analytics Engine cluster is built on modern virtualization and container technology. The Hadoop or Spark distribution that runs on each node is encapsulated in a Docker container, which can be initialized or shut down in a matter of seconds.

Maintainability

Since the clusters only exist while they are being used, it is also much easier to perform maintenance. If the user wants to move to newer versions of Hadoop or Spark, they can simply create a new cluster with the new version of the software. The upgrade has no impact on existing clusters or other users' environments, and there is no risk of data loss or corruption because the master copy of the data is stored in the object store, not in the cluster.

Openness

IBM Analytics Engine harnesses Hortonworks Data Platform (HDP), an ODPi-compliant, open source Apache Hadoop distribution. Unlike many other Hadoop distributions, which include proprietary software, HDP is 100 percent open source. By utilizing this open source platform, customers can extend their existing open source investments and benefit from the contributions of the open source community.

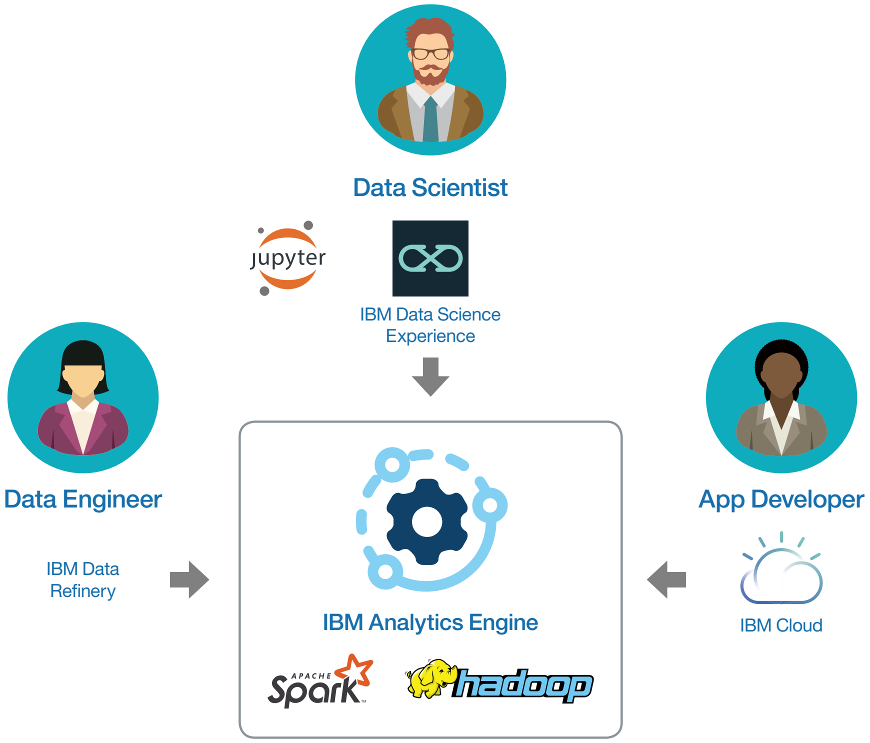

Figure 1. IBM Analytics Engine, built on Spark and Hadoop, provides an environment for deploying analytics applications and a foundation for IBM Watson Data Platform services.

Ease of use

IBM Analytics Engine provides a single, central point of control for creating and managing both Hadoop and Spark environments in the cloud. As a result, data scientists no longer need to spend time procuring, configuring and managing servers, installing and updating software, or troubleshooting problems and removing bottlenecks. Instead of worrying about the infrastructure, they can focus entirely on their real job: exploring data, building models, and analyzing results.

Users can interact with IBM Analytics Engine directly via an intuitive, browser-based web user interface, or programmatically via a command-line interface or REST API. They can define, customize and create Hadoop or Spark clusters in minutes, and can even install third-party libraries and configurations across the whole cluster with a few clicks.

End-to-end integration

IBM Analytics Engine is part of IBM Watson Data Platform, which means it integrates seamlessly with other services such as:

- IBM Data Catalog, for metadata management and cataloging
- IBM Data Science Experience, for data exploration
- IBM Data Refinery, for integrating, cleaning, enriching and refining large data sets
- IBM Watson Machine Learning, for training and testing predictive models and neural networks

Instead of each data science team using different tools, Watson Data Platform provides a standard set of best-of-breed solutions to support every step in the data science value chain.

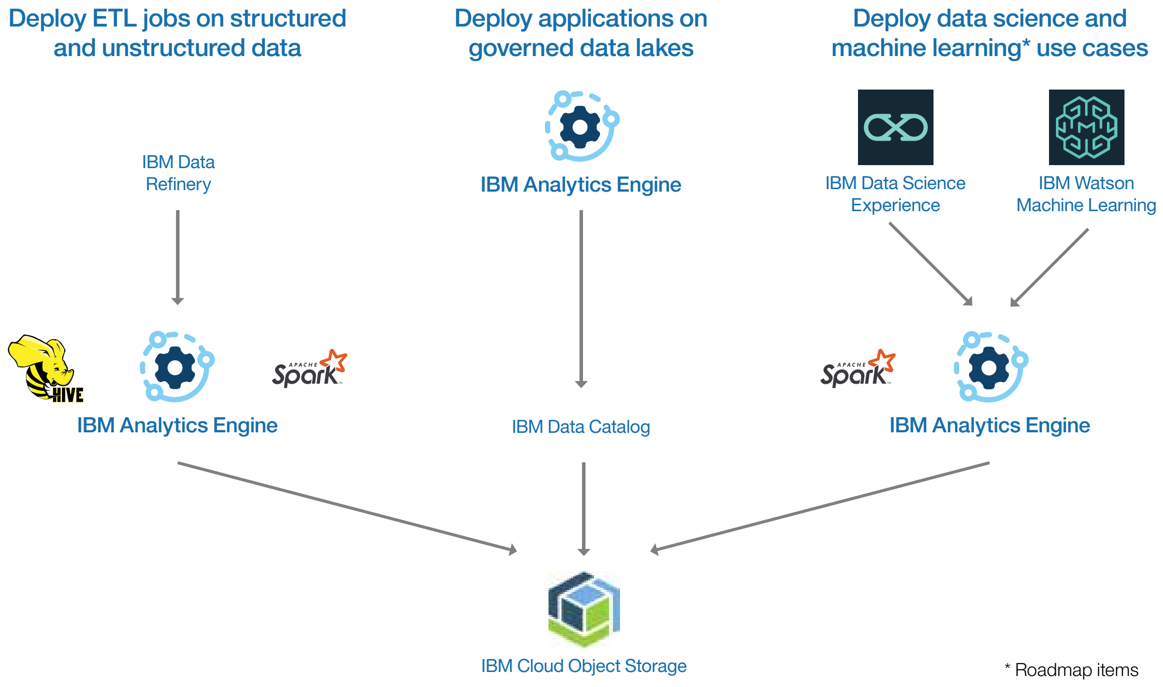
Deploy ETL jobs on structured and unstructured data
Deploy applications on governed data lakes
Deploy data science and machine learning* use cases
IBM Data Refinery
IBM Analytics Engine
IBM Data Science Experience
IBM Watson Machine Learning
HIVE
Apache Spark
IBM Analytics Engine
IBM Data Catalog
Apache Spark
IBM Analytics Engine
IBM Cloud Object Storage
* Roadmap items

Figure 2. IBM Analytics Engine provides support for newer types of workloads and traditional Hadoop use cases in an efficient manner through IBM Watson Data Platform.

Moreover, as each of the Watson Data Platform services is built on a shared architecture of microservices, integration between the different components is seamless and consistent. This eliminates the need to write custom scripts to link tools together, and makes the flow of data more reliable and easier to maintain.

Use cases for IBM Analytics Engine

IBM Analytics Engine is designed as a multi-purpose big data processing engine that can be leveraged by other solutions and tools in many different use cases. Here are just a few examples of how IBM Analytics Engine can help solve real-world big data challenges.

1. Simplifying data governance

Traditional data lakes built on Hadoop often suffer from data governance issues. To restrict access to sensitive data, the Hadoop cluster needs to be segregated into different zones for different sets of users, which requires a large amount of configuration and maintenance.

This problem can be solved by combining IBM Analytics Engine with IBM Cloud Object Storage and IBM Data Catalog. Instead of building the data lake as a single, persistent Hadoop cluster that has to handle all the data, governance configurations and computation work, you separate these three concerns.

All data assets are stored permanently in IBM Cloud Object Storage and can only be accessed via IBM Data Catalog. The catalog contains rich metadata about each of the data assets: their lineage, what kind of information they contain, and which users are allowed to access them.

When an authorized user finds the dataset they need, they can then use IBM Analytics Engine to spin up a temporary cluster and load the data into it for analysis. Once the analysis is complete, the results can be saved back into object storage and automatically added to the catalog for other users to find. Once the cluster has served its purpose, it can be deprovisioned.

Throughout the process, governance rules are applied automatically, and there is no need for complex cluster configuration—saving time for data stewards and IT teams, and making it easier for users to find the data they need.

2. Reducing the cost of disaster recovery

Disaster recovery has been a significant pain point for traditional Hadoop-based data lakes. In most cases, the only viable solution has been to set up a secondary cluster and implement expensive proprietary software to keep it in sync with the main cluster.

The combination of IBM Analytics Engine and IBM Cloud Object Storage makes disaster recovery both more convenient and less expensive. The object storage solution automatically replicates multiple copies of data between nodes, and it can even be set to keep copies of data in different geographical regions, avoiding the risk of data loss even if a whole data center goes offline.

In the event of such an outage, all a user needs to do is pass their configuration file to IBM Analytics Engine and spin up a new cluster with identical settings to the old one. This new cluster will connect automatically to the object storage environment and start working on one of the remaining copies of the data.

This simple architecture significantly reduces the amount of effort required to set up a disaster recovery environment and makes it quicker and easier to recover if a disaster occurs.

3. Streamlining data science and machine learning workflows

Data science workflows depend on the ability to explore large data sets and train and test predictive models and neural networks efficiently—yet currently, data scientists spend much of their time writing scripts to move data between environments, or configuring and maintaining Hadoop and Spark clusters.

Tools such as IBM Data Science Experience and IBM Watson Machine Learning leverage IBM Analytics Engine to make these processes more seamless and automated. For example, a data scientist who wants to explore a large data set can simply open a notebook in IBM Data Science Experience, associate an IBM Analytics Engine Spark cluster with it, and start looking for insights straight away.

Similarly, when a data scientist is ready to begin training their models, they can use IBM Watson Machine Learning to invoke IBM Analytics Engine to set up a cluster and launch the job within minutes.

Conclusion

IBM Analytics Engine makes it easier for data scientists, data engineers and developers to develop and deploy analytics applications. Users can spin up Hadoop and Spark clusters within minutes and manage them from a single point of control. As a result, users no longer need to worry about the underlying infrastructure and can focus on their core role of delivering insight to the business.

Under the hood, the separation of compute from storage allows companies to scale their big data landscape in line with business requirements and to avoid investing in infrastructure that they don't need. The use of object storage helps to protect data and improve availability. Meanwhile, the fact that clusters can be spun up and spun down dynamically as needed helps to simplify maintenance and upgrades.

Finally, IBM Analytics Engine integrates seamlessly with other IBM Watson Data Platform services, opening up new possibilities for end-to-end data science workflows, improving collaboration, and eliminating the complexity of fragmented tool-chains.

For more information

To learn more about IBM Analytics Engine, contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/analytics/us/en/watson-data-platform/analytics-engine/

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
November 2017

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Please Recycle