Choose IBM Open Platform for your Hadoop and Spark projects

Explore this mature Apache Hadoop and Apache Spark distribution

Andrea Braida

November 08, 2016
(First published November 02, 2016)

Learn about the IBM Open Platform (IOP), which includes a set of open-source components that are supported by IBM’s deep Hadoop, Spark, and big data expertise. IOP includes what we believe is the best combination of components to deliver a comprehensive range of capabilities across the most common big data use cases. Find out about the components and their value.

When Apache Hadoop was launched in 2006, it was one of the most ground-breaking and disruptive technologies in the history of IT. Today, as more organizations recognize the value of big data, Hadoop’s relevance for mainstream businesses has never been higher. With the emergence of Apache Spark and other new open source projects, the Hadoop ecosystem now offers faster and richer analytics tools to help companies exploit big data.

The IBM Open Platform (IOP) includes a set of open source components that are supported by IBM’s deep Hadoop, Spark, and big data expertise. IOP is ODPi compliant and includes what we believe is the best combination of components to deliver a comprehensive range of capabilities across the most common big data use cases.

This article describes the distribution and several of its IOP components. In follow-on articles, we’ll provide more in-depth discussion from IBM subject matter experts about specific Apache Hadoop projects and related use cases.

Exploring the components of the distribution

Some of the functions and their components of the IOP distribution, as shown in Figure 1, include:

- Processing: Spark and MapReduce
- Integration: Sqoop, Flume, and Kafka
- Storage: HDFS
- Security: Ranger and Knox
- Scripts: Pig
- Search: Solr
• Management: Ambari, YARN, Oozie, and Slidr
• Data science: System ML, Hydra R, Spark R, and Titan
• SQL and no SQL: Phoenix, HBase, and Hive

Figure 1 graphically shows the current components and how they interrelate.

**Figure 1. IBM Open Platform: Hadoop and Spark distribution**
Learning more about IOP components

This section describes some of the IOP components and their value in processing data.

Apache Spark for speed and flexibility

For accelerated processing time and for deeper and richer analytics, the IBM Open Platform includes Apache Spark, which is arguably the most talked about technology in the current ecosystem, and one that many people believe is the future of Hadoop. Spark is an open source, in-memory compute engine and analytics platform that was designed to make big data analytics and development easier. Spark is best known for its speed (in-memory processing provides up to 100 times faster processing speed than other big data technologies), ease of use, and flexibility to run in a variety of programming environments. Spark effectively diminishes silos and empowers anyone to access data from anywhere, while reducing the number of tools required to do it. It supports a new generation of data applications that are built to harness the Internet of Things and whatever comes next.

Spark accelerates the delivery of results by enabling analytics and other complex algorithmic transformations to be performed in memory without writing the results of each intermediate job to disk. This makes it possible to do the following:

- Speed up batch and ETL processes
- Perform near-real-time analytics using micro-batch streaming on a multitude of data sources
- Leverage built-in machine learning libraries to create predictive models
- Perform SQL queries on unstructured data

Apache Hive query language for long-term analytics

For longer-running, less time-sensitive analytics use cases, Apache Hive reduces the complexity of interacting with Hadoop. Hive provides an intuitive declarative query language that reduces the need to write MapReduce jobs, which makes Hive more tractable for data scientists who aren't Java programmers. Hive is an ideal option for queries that analyze large data sets collected over a long period of time, such as calculating trends or creating summaries.

Apache HBase for fast queries

For time-sensitive queries, transactions, and jobs that involve writing data as well as reading it, Apache HBase provides a highly scalable approach, holding data in a NoSQL store based on key-value pairs that can be accessed and written very quickly even on the largest data sets.

Apache Phoenix for an SQL front-end for HBase

Apache Phoenix adds a high-performance SQL front-end for HBase, opening it up to a much wider audience of programmers and data scientists.

Titan for analyzing complex networks

Titan enables highly efficient analysis of relationships between nodes in complex networks. This is a key challenge when thousands or even millions of people or devices share various types of connections with each other, including social networking and telecommunications use cases.
Apache Oozie for scheduling

Apache Oozie is a scheduler to trigger Hadoop jobs to be run at specific times or under specific conditions, such as whenever new data becomes available.

Apache Flume for feeding streams of data

Apache Flume provides seamless integration capabilities by enabling you to create a pipeline to feed streams of data into or out of your Hadoop cluster.

IBM BigInsights for scaling applications

IBM's Hadoop offerings, available on premises and on the cloud, provide additional enterprise-grade capabilities that help to scale analytics and applications quickly and easily. IBM is continuously developing these capabilities. The latest version, IBM BigInsights® 4.2, was released in June of 2016.

Conclusion

Selecting the right distribution is key to having the flexibility and support that you need to build a robust analytics strategy. The list of components included in the IBM Open Platform distribution continues to evolve in order to include the very latest useful open-source technologies. To learn more about the IBM Open Platform and IBM's other Hadoop offerings, visit ibm.com/analytics/us/en/technology/biginsights.
Related topics

- Hadoop offerings from IBM
- IBM BigInsights solution brief
- Overview of SSL support in IOP 4.2

© Copyright IBM Corporation 2016
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)