GridGain

Built On Apache Ignite

APPLICATION NOTE

Apache Ignite and GridGain: In-Memory Data Management and Acceleration for Apache Spark

Apache® Spark™ is a leading fast and general purpose open source engine for large scale data processing of streaming data. It has become the de facto processing engine for Apache Hadoop™. But while Spark supplanted MapReduce for processing, no real-time data management solution has emerged. Spark doesn't manage data or track state. As a result, developers using Spark often write extensive code to ingest, prepare, enrich, store, manage data and state.

Apache® Ignite™ and GridGain® provide the most extensive in-memory data management and acceleration platform for Spark. Ignite is an open source platform that provides an in-memory data grid (IMDG), in-memory database (IMDB), streaming analytics, and continuous learning framework for machine and deep learning. GridGain is the leading in-memory computing platform for real-time business and includes enterprise-grade security, deployment, management and monitoring features which are not in Ignite. GridGain Systems contributed the code that became Ignite to the Apache Software Foundation and continues to be the project's lead contributor.

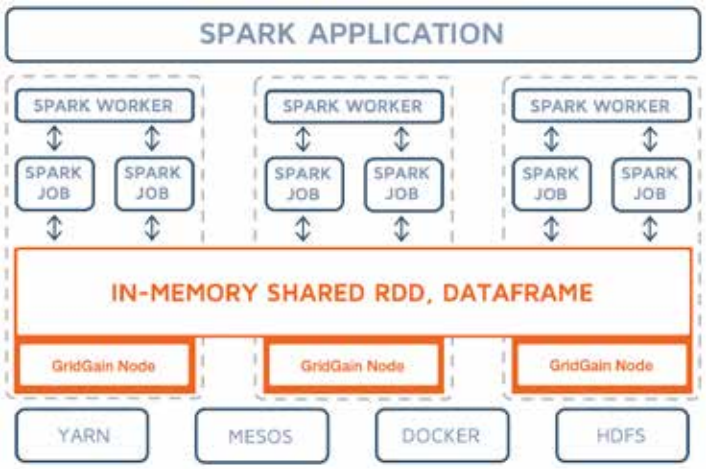
Ignite provides the ideal underlying in-memory data management technology for Apache Spark with in-memory support for both stored "data at rest" and streaming "data in motion." It makes many Spark tasks simple, including stream ingestion, data preparation and storage, stream processing, state management, streaming analytics, and machine and deep learning.

IGNITE IN-MEMORY DATA MANAGEMENT

Ignite adds integrated data management with in-memory speed and unlimited scale for any data used by Spark. Ignite is a distributed in-memory SQL and key-value store which can sit on top of all popular databases. It can store data as a distributed IMDB that outperforms traditional databases on streaming data ingestion and ACID transactions. On top of databases, Ignite merges all the structured, semi-structured and unstructured data together into a single in-memory data access layer. The combined data is accessible using ANSI-99 compliant SQL or key-value APIs across different programming languages as well as Spark RDDs, DataFrames and HDFS. By moving all data into RAM, Ignite eliminates the delays of disk-based databases.

Ignite also eliminates another major bottleneck with data: the network. Spark jobs and other computing can be collocated on the same machine with any required data when using Ignite. This eliminates the delay of moving data over the network for processing. Ignite allows nodes of a cluster to be installed on every machine running Spark jobs and uses configurable data affinity to distribute the right data to the right nodes. Developers can also leverage the Ignite Compute Grid for real-time analytics, high performance computing, microservices, or machine or deep learning. Ignite distributes any custom Java, .NET or C++ code across the cluster and executes it locally on each node using massively parallel processing (MPP). Ignite provides a broad, integrated implementation of distributed MPP algorithms including distributed SQL and machine and deep learning.

Spark performs stream processing and Ignite provides data locally in memory using Spark APIs. This makes big data analytics and in-place continual learning for real-time responsiveness and automation a reality.



STREAM INGESTION AND DATA PREPARATION

Ignite is the ideal stream ingestion and data preparation engine for Spark. It can ingest, store, process, analyze and publish large volumes of streaming data with low latency, unlimited scalability and high availability. Ignite can then merge streaming and stored data and expose it as RDDs or DataFrames in memory, in real-time. Ignite can ingest

millions of events per second on a moderately-sized cluster. Ignite is used with major streaming technologies including Apache Camel™, Apache Flink™, Apache Flume™, Apache Kafka™, Apache RocketMQ™, Apache Spark, Apache Storm™, Java Message Service (JMS), MQTT, Twitter® and ZeroMQ to ingest, process and publish streaming data. Once loaded into the cluster, developers can leverage built-in Ignite libraries for concurrent data processing, including distributed SQL queries and machine and deep learning. Clients can also subscribe to continuous queries which execute and identify important events as streams are processed.

SPARK RDD AND DATAFRAME SUPPORT

Ignite provides native support for Spark RDDs and DataFrames. The Ignite RDD API allows Ignite caches to be collocated on the same machine and accessible directly inside Spark processes that are executing Spark jobs. Unlike Spark RDDs, which are immutable or read-only, Ignite RDDs are mutable. Spark developers can directly read from and write to Ignite to save data as well as share data and state across Spark jobs. Ignite also extends the DataFrame API to allow developers to use Ignite for data storage. Whenever data is updated, it writes directly to Ignite and is immediately accessible to others.

SPARK SQL ACCELERATION

Apache Spark supports a fairly rich SQL syntax but it doesn't support data indexing so Spark must do full scans all the time. Spark queries may take minutes, even on moderately sized data sets. Ignite optimizes Spark's query execution plans to leverage Ignite's distributed SQL and advanced indexing. Ignite also optimizes underlying Spark SQL plans within DataFrames to use Ignite SQL when appropriate. Spark can take advantage of the advanced indexing and massively parallel processing (MPP) distributed joins in Ignite to improve Spark SQL query performance by up to 1000x.

HDFS ACCELERATION

It is possible to share state between Spark jobs and applications using the Apache Ignite In-Memory File System (IGFS) with files instead of RDDs. IGFS implements the Hadoop File System (HDFS) API and an in-memory MapReduce implementation optimized for MPP on the Ignite Compute Grid. Ignite can be used as a native Hadoop file system, just like HDFS. Ignite plugs in natively to any Hadoop and Spark environment and can be used in place of HDFS in a plug-n-play fashion.

MACHINE AND DEEP LEARNING

Spark plus Ignite for end-to-end in-memory computing enables companies to rapidly deliver new in-process hybrid transactional/analytical processing (HTAP) applications. With Spark and Ignite, companies can train models against massive data sets and re-run training in "midstream" during Spark processing to improve the models based on the latest data. Ignite provides several standard machine learning algorithms optimized for MPP including linear and multi-linear regression, k-means clustering, decision trees, k-NN classification and regression. It also includes a multilayer perceptron and TensorFlow integration for deep learning. Developers can develop and deploy their own algorithms across any cluster as well using the Compute Grid.

Contact GridGain Systems

To learn more about how GridGain can help your business, please email our sales team at sales@gridgain.com, call us at +1 (650) 241-2281 (US) or +44 (0)208 610 0666 (Europe), or go to www.gridgain.com/contact to have us contact you.

About GridGain Systems

GridGain Systems is revolutionizing real-time data access and processing with the GridGain in-memory computing platform built on Apache® Ignite™. GridGain and Apache Ignite are used by tens of thousands of global enterprises in financial services, fintech, software, e-commerce, retail, online business services, healthcare, telecom and other major sectors, with a client list that includes ING, Raymond James, American Express, Societe Generale, Finastra, IHS Markit, ServiceNow, Marketo, RingCentral, American Airlines, Agilent, and UnitedHealthcare. GridGain delivers unprecedented speed and massive scalability to both legacy and greenfield applications. Deployed on a distributed cluster of commodity servers, GridGain software can reside between the application and data layers (RDBMS, NoSQL and Apache® Hadoop®), requiring no rip-and-replace of the existing databases, or it can be deployed as an in-memory transactional SQL database. GridGain is the most comprehensive in-memory computing platform for high-volume ACID transactions, real-time analytics, web-scale applications, continuous learning and hybrid transactional/analytical processing (HTAP). For more information on GridGain products and services, visit www.gridgain.com.