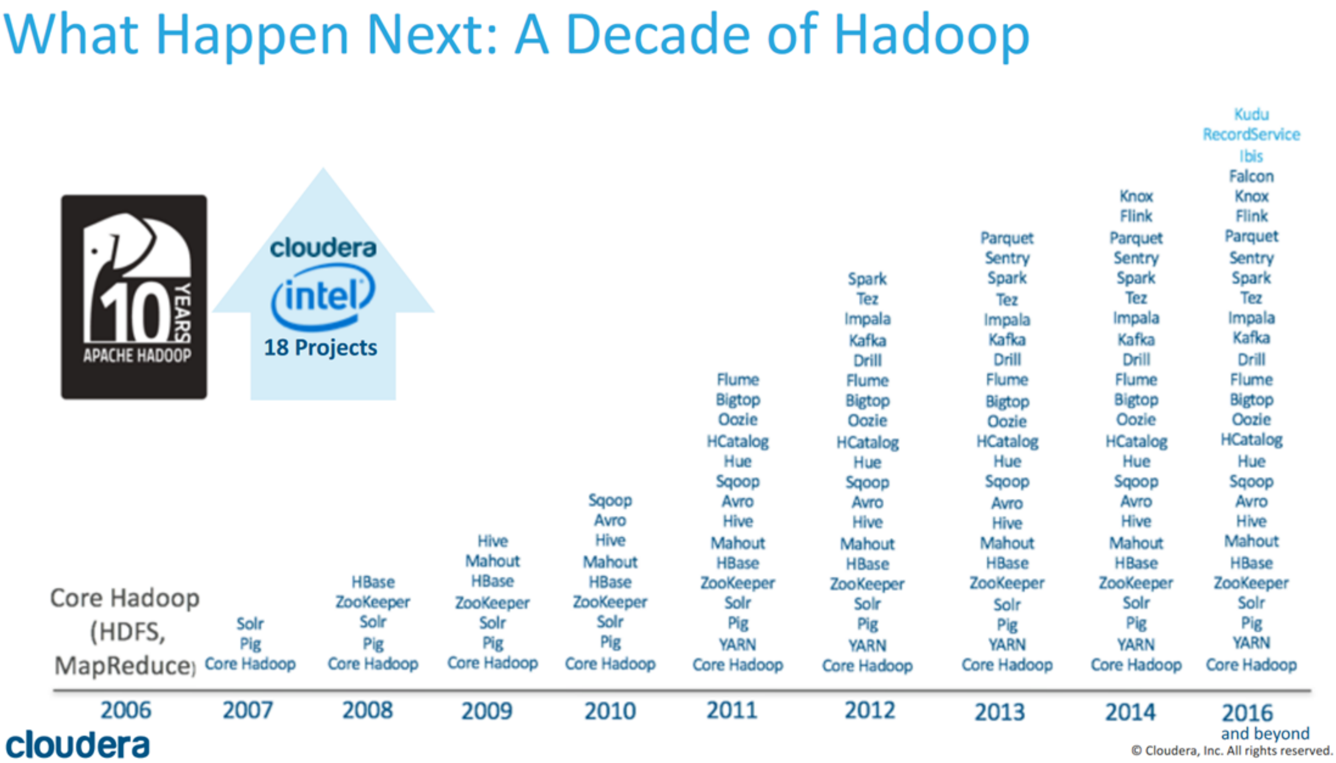
Article
Yuri Marx · Oct 19, 2021 5m read

Open Exchange

Big Data components and the InterSystems IRIS

In the last years the data architecture and platforms focused into Big Data repositories and how toprocess it to deliver business value. From this effort many technologies were created to process tera and petabytes of data, see:



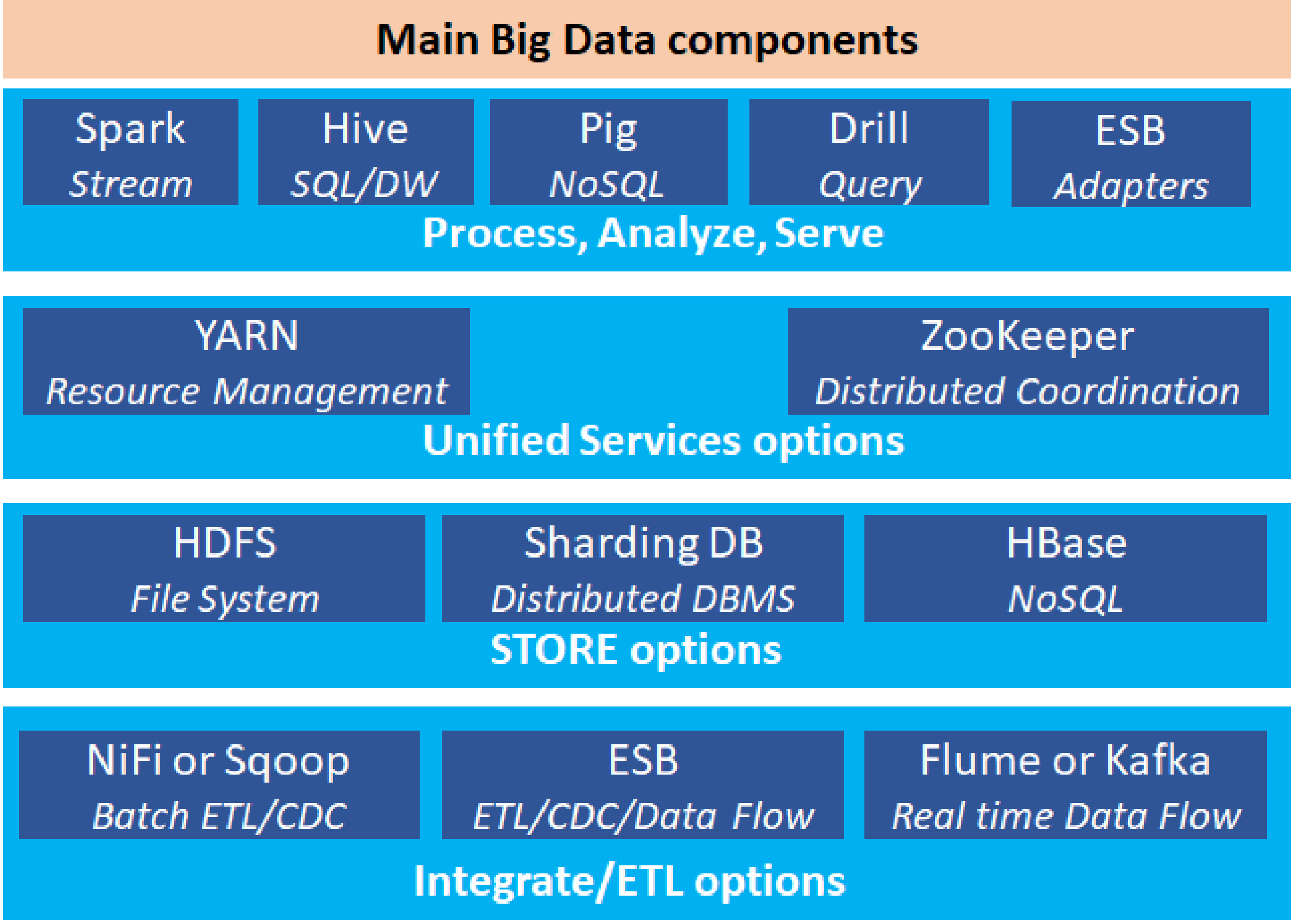
What Happen Next: A Decade of Hadoop

2006	2007	2008	2009	2010	2011	2012	2013	2014	2016 and beyond
Core Hadoop (HDFS, MapReduce)	Solr Pig Core Hadoop	HBase ZooKeeper Solr Pig Core Hadoop	Hive Mahout HBase ZooKeeper Solr Pig Core Hadoop	Sqoop Avro Hive Mahout HBase ZooKeeper Solr Pig Core Hadoop	Flume Bigtop Oozie HCatalog Hue Sqoop Avro Hive Mahout HBase ZooKeeper Solr Pig YARN Core Hadoop	Spark Tez Impala Kafka Drill Flume Bigtop Oozie HCatalog Hue Sqoop Avro Hive Mahout HBase ZooKeeper Solr Pig YARN Core Hadoop	Parquet Sentry Spark Tez Impala Kafka Drill Flume Bigtop Oozie HCatalog Hue Sqoop Avro Hive Mahout HBase ZooKeeper Solr Pig YARN Core Hadoop	Knox Flink Parquet Sentry Spark Tez Impala Kafka Drill Flume Bigtop Oozie HCatalog Hue Sqoop Avro Hive Mahout HBase ZooKeeper Solr Pig YARN Core Hadoop	Kudu RecordService Ibis Falcon Knox Flink Parquet Sentry Spark Tez Impala Kafka Drill Flume Bigtop Oozie HCatalog Hue Sqoop Avro Hive Mahout HBase ZooKeeper Solr Pig YARN Core Hadoop

cloudera

© Cloudera, Inc. All rights reserved.

The fundamental piece to the Big Data technologies is HDFS (Hadoop Distributed File System). It is a distributed file system to store tera or petabytes of data into arrays of storages, memory and CPU working together. In addition of the Hadoop we have other components, see:



In a Big Data project the first step is create the data ingestion, integration, transformation/enrichment, with data quality as a requirement. In this area we have:

- Apache Kafka: data event flow, in real time asyncronous events, were each event deliver data items to be ingested by the Big Data using Kafka clients. It is possible process milions of data event per second, creating real time data streams.
- Apache NiFi: it is an ETL capable to create integration maps that connect with multiples formats, technologies and type of data sources to deliver data to HDFS, Databases or HBase repositories. Acts also as a CDC tool.
- Apache Sqoop: it is an Data Ingestion tool for SQL Databases.
- Apache Flume: collect data from the log of the datasources to push to the HDFS or HBase.
- ESB: it is a service bus used to connect internal and external datasources into a corporate data backbone to send, receive and process data between enterprise data resources, including Big Data repositories. The ESB has connectors/adapters to the main data resources, like Kafka, SQL Databases, NoSQL databases, E-mail, Files, FTP, JMS messages and other. The InterSystems IRIS has an ESB. The ESB can automate data process, delivering complex integrations to the Big Data repositories.
- HDFS: it is the most used resource to store big data volumes, with performance, realiability and availability. The data storage is distributed into data nodes working in an active-active HA and using common machines. So, it is cheaper than store into RDMS products.
- HBase: it is like HDFS, but it is used to store NoSQL data, like objects or documents.
- Sharding DB: are data stores that use distributed data nodes to allow process Big Data SQL or NoSQL repositories into proprietary formats. The main example is MongoDB, but InterSystems IRIS it is a Shard DB too, including options to store SQL, NoSQL (JSON) or OLAP datastores.
- YARN and ZooKeeper: are tools to centralize and allows the Big Data tools share configuration, resources, resource names and metadata.
- Apache Spark: Apache Spark is a distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It

provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing (AWS definition). The InterSystems IRIS has a Spark adapter that can be used to read and write HDFS data.

- Apache Hive: it is a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale. A data warehouse provides a central store of information that can easily be analyzed to make informed, data driven decisions. Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data. What makes Hive unique is the ability to query large datasets, leveraging Apache Tez or MapReduce, with a SQL-like interface (AWS definition). The InterSystems IRIS don't have native adapter for Hive, but I created it (see: https://community.intersystems.com/post/using-sql-apache-hive-hadoop-big-data-repositories).
- Apache Pig: it is a library that runs on top of Hadoop, providing a scripting language that you can use to transform large data sets without having to write complex code in a lower level computer language like Java. The library takes SQL-like commands written in a language called Pig Latin and converts those commands into Tez jobs based on directed acyclic graphs (DAGs) or MapReduce programs. Pig works with structured and unstructured data in a variety of formats. (AWS definition).
- Apache Drill: it is a tool to do queries for the main Big Data products using SQL sintax. Drill supports a variety of NoSQL databases and file systems, including HBase, MongoDB, MapR-DB, HDFS, MapR-FS, Amazon S3, Azure Blob Storage, Google Cloud Storage, Swift, NAS and local files. A single query can join data from multiple datastores. For example, you can join a user profile collection in MongoDB with a directory of event logs in Hadoop (Apache Drill definition). I will build an adapter to Apache Drill in the near future.
- ESB Adapters: it is an interoperability platform, with ready connectors to query and process data from multiple formats and protocols. The InterSystems IRIS has adapter to HTTP, FTP, File, SQL, MQTT (IoT) and many other datasources. These adapters can compose data flows (productions) to deliver the data for multiple targets, including hadoop repositories.

InterSystems IRIS it is a good Big Data architecture option with:

1. An excellent integration/ETL layer, as an ESB platform.
2. A fantastic distributed data store when using IRIS Database Sharding.
3. An important tool to query and process Big Data information, using its adapters. The InterSystems IRIS can also deliver reports, BI Dashboards and data microservices.
4. A good tool to work with Kafka, Spark, Hadoop and Hive, to deliver Big Data with additional resources, like ESB, Analytics, Machine Learning, Data workflows (BPL productions) and native support to the main languages (Python, R, Java, .Net, Node.js).

See a sample of this into my new app:
https://openexchange.intersystems.com/package/IRIS-Big-Data-SQL-Adapter. Thanks! And enjoy!

#Big Data #InterSystems IRIS #InterSystems IRIS for Health
Check the related application on InterSystems Open Exchange

Source URL:https://community.intersystems.com/post/big-data-components-and-intersystems-iris