kognitio

Kognitio Analytical Platform

Analytics for the Data Lake

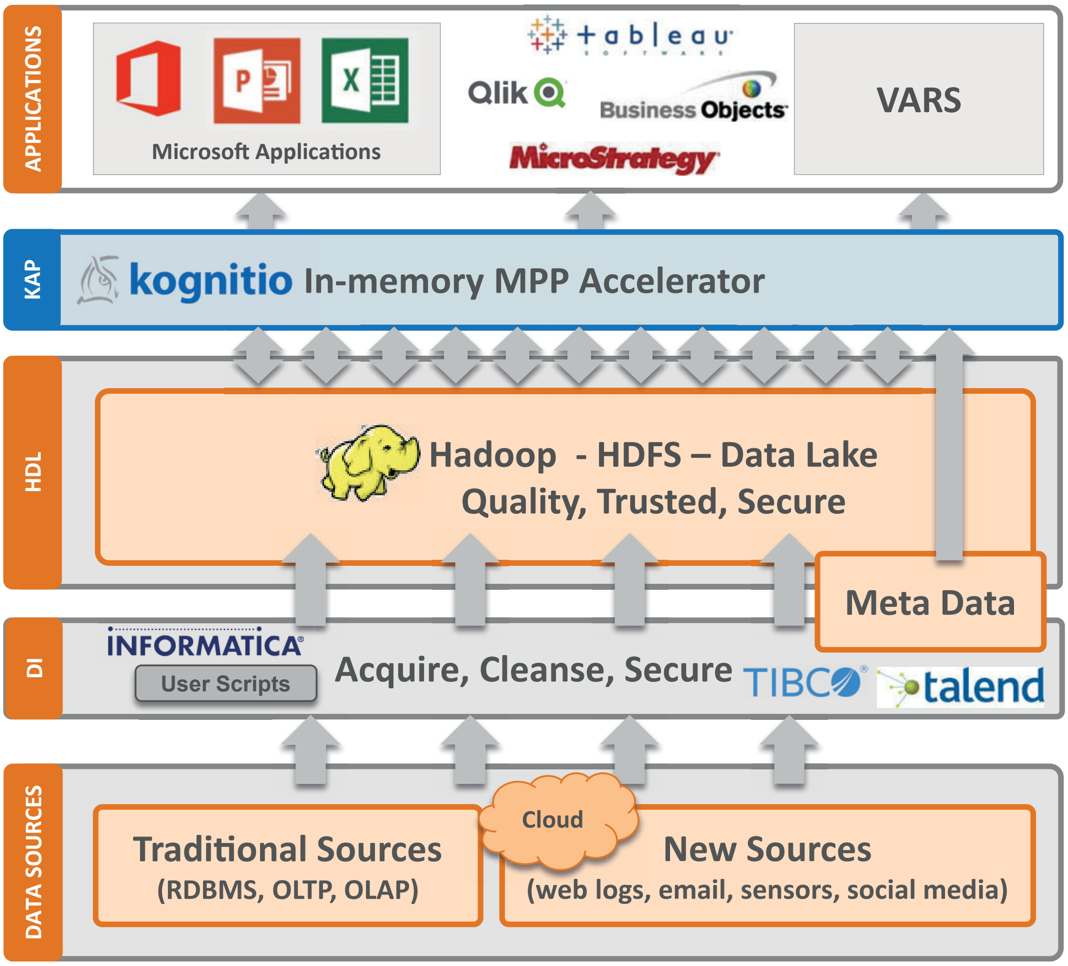
Kognitio is a pioneer in high-performance, scalable Big Data analytics for Data Science & Business Intelligence

Advanced Analytics Platform for Hadoop Data Lake

Introduction

The purpose of this paper is to articulate how the Kognitio Analytical Platform can dramatically add business value to a ***Hadoop Data Lake*** (HDL) implementation.

Hadoop (HDFS File System) is becoming the de-facto repository for Data Lakes. By selecting Hadoop as the persistence engine, it means that customers can have a truly scalable repository for all of their new Big Data requirements. In addition, the repository can also store the metadata, keys and pointers to the wide range of operational data which may be stored in business applications, data warehouses, files and legacy systems. HDL is therefore the beneficiary recipient of all of the cleansed, trusted, secure quality data that is available.



Kognitio has scalable high-performance parallel data load integration with HDFS and Map-Reduce which pulls full-volume Hadoop data via simple mappings, and streams it into the in-memory ***Kognitio Analytical Platform*** (KAP) at speeds which currently approach 15 terabytes an hour. This is around two orders of magnitude faster than standard open source products (e.g. HIVE). This means that it can pull extremely large data sets into memory, process and analyse them, and then, additionally, write the results back into Hadoop at the same blistering speeds. The massively parallel architecture of KAP combined with tight integration with Hadoop enables a wide range of analytic business applications which, by definition, require a combination of huge data sets, intense compute capability and fast response times.

Business Proposition

The world of data is expanding exponentially – operational transactions, data warehouses, cloud data and interaction data from social media, web logs and machines. The data integration industry has a well proven process to accept data feeds from all of these sources and integrate them into a coherent information base. Master data management, data quality and security products add a further level of integrity resulting in a reliable, trusted base of information on which to base informed business decisions.

There is a hierarchy of complexity and scale within the world of Analytics – from simple reporting; through SQL based Business Intelligence; and up to data science led applications (both SQL and No-SQL based) for solving complex problems. Often this moves the focus from retrospective analysis of historical data to predictive analysis of future events - to anticipate – looking out of the front window rather than the rear window view of the customer business.

Kognitio adds a scalable, high performance, in-memory platform with associated massively parallel computational capability to enable the Hadoop Data Lake to participate fully in this new world of next-generation analytics. This participation is about making users interactions with big data easy and allowing them to use the applications interfaces and tools that they have engaged with in the BI revolution. Plugging some BI and analytics tools straight into Hadoop can be a frustrating and seriously underwhelming experience that slows the potential innovation from the wealth of data available. The sweet spot is the top half of the Analytics hierarchy. The Kognitio Analytical Platform provides a cast iron and demonstrable response to the traditional Hadoop questions about maturity, performance, reliability, throughput and usability for modern interactive analytics.

Product

The Kognitio Analytical Platform is the glue that connects mass data stored in Hadoop with end-user business applications, analytical models and tools. It is a scalable, in-memory, parallel analytics engine which can take very large data sets into memory at two orders of magnitude faster than open source products (e.g. HIVE), process and analyse them, and then write the results back into Hadoop at the same performance levels.

To applications and tools, KAP looks like any other standard SQL database or MDX cube, but under the covers it is very different. KAP is not a storage or data filtering engine, but instead has been designed to focus on the compute intensive steps required for complex data analytics. It does this by automatically and efficiently parallelising all operations, across every CPU core, in every server made available to it. To ensure CPU cores are never waiting for disk IO, the data of interest is held in memory and all temporary result sets are kept only in memory. By scaling from one server to hundreds of multi-core servers, KAP can enable the timely execution of very complex operations, across even the largest of data sets.

Hadoop has the capability of allowing customers to store very large volumes of persistent data – often running into petabytes. To gain access to such large data sets Kognitio provides a high performance fully integrated connector based on standard APIs which can bring the data into memory at very high speeds. This loader technology operates at the speed at which data can be loaded into memory, which in turn is a function of the size of both the KAP and Hadoop cluster, but on a moderately sized platform Kognitio has demonstrated load rates of 15 terabytes an hour. Once the data has been analysed, or the models run, KAP can write any result sets, insights and updated models back into Hadoop at the same high performance levels.

As well as having very rich SQL and MDX support, KAP has the ability to embed any binary or script into the queries, enabling massively parallel execution of NoSQL functions. The binary or script can be written, in-line, in virtually any language, R, Python, Java, C etc. Any code can be encapsulated into callable routines or embedded in views so that users can interact with on-the-fly complex calculations via their standard tools - a report column could now be a dynamically calculated forecast. These features combine to create a very flexible platform for large scale data analytics without the need for complex engineering or teams of programmers.

To complete the integration picture, KAP also has seamless connectivity to a range of other data storage systems e.g. Oracle, SQL Server, Netezza, Teradata etc. and to cloud storage services such as AWS S3.

The Need

The market positioning is very compelling. The business imperative is for customers to have access to sufficient relevant information to improve the quality of their decision making. The IT initiative is to provide the platform on which the customers can run their demanding analytics. The business wants to continue with its investments in BI tools and analytical software and not have to start all over again. Hadoop has become the de-facto standard for storing massive sets of data on low cost hardware, which enables customers to store full volumes of transaction and interaction data at costs which were previously prohibitive. This Hadoop Data Lake means that customers can have a truly scalable, near infinite, repository for all of their new Big Data requirements. In addition, the repository can also store the metadata, keys and pointers to the wide range of operational data which may be stored in business applications, data warehouses, files and legacy systems. HDL is therefore the beneficiary recipient of all of the cleansed, trusted, secure quality data that is available.

This empowered analytics capability can support a wide range of powerful solutions across all industries. Examples include: tariff and price modelling for Telcos using full volume call detail records; demand forecasting and supply chain optimisation for retailers by analysis of shopping cart records; pay-as-you-drive insurance pricing using GPS location tracking and telematics for analysing driver behaviour; credit card transaction analysis for behaviour scoring to control risk and identify fraud patterns; and topological applications for analysing satellite images for governments to track fishing boat movements etc.

Customers will have the choice of using a wide variety of end-user tools to query the HDL and to produce simple reports and analytics. Tableau is the current leader, while many companies still use traditional tools such as Business Objects, Cognos and Microstrategy. However, Hadoop does have some well-recognised limitations in terms of latency and performance, and the usability of any interactive analytics/BI tool is highly dependent on the performance of the underlying data source. The benefits of Hadoop as the preferred data repository can often be soured by performance and usefulness issues for analytics tools vendors as customers resort to hand-coding to try to resolve these issues. The problem compounds when customers have the requirement for more complex or compute intensive reporting and analytics, and the market is moving steadily towards data science-led applications.

The Kognitio Analytical Platform solves these major performance issues, and their resultant usefulness, by quickly pulling the required data model into memory and then employing parallel computation technology to enable the selected analytics or BI tool to perform effectively for the end-user analyst. Any data model can be quickly acquired from the lake, refined and embellished on-the-go, and the results rapidly interpreted, supporting the workbench style of use preferred by many analysts and data scientists. The Kognitio difference here will be the scale of data available, the speed of interaction and operation, and scale of data being analysed.

Use Cases

Firstly, enterprises with a commitment to Big Data and with Big Data or Data Science departments already in place.

A leading Credit Card supplier is a very good example of such a customer, where a five terabyte RAM Kognitio Analytical Platform was deployed to gain insight into a nine petabyte lake of event and interaction data. The marketing analytics team scoured the market to allow their investment in tools such as Tableau and Microstrategy to exploit the Hadoop store without hampering the end-user experience - Kognitio was the only viable option for robust, stable, timely analytics.

Another example is a US retailer who allows a mixed workload of 4000 queries an hour to be 'thrown' at the system with 97% of them typically responding in under 2 seconds! Their Hadoop cluster is being built alongside and will represent a deeper store of raw data to be exploited as a future data lake.

Secondly, applications partners (VARS). There are several examples of analytical application VARS in different industries. One such VAR is a Global Loyalty Management company, who provide a Kognitio-based analytics solution based on full-volume shopping basket (SKU-level) data to retailers across the world. They now have successful customers across North America, Europe and Asia.

Summary

Kognitio has a scalable, high performance solution which enables customers across many industries to gain valuable insight into their businesses by analysing and modelling very large data sets derived from both traditional and new data sources. Kognitio provides a simple bridge to the data lake allowing the traditional tools and applications to plug-and-play while giving them high-performance queries and fast response times that the end-users love. The data lake only delivers business value if it is actively and continuously used by inquisitive users.

www.facebook.com/kognitio

www.twitter.com/kognitio

www.linkedin.com/company/kognitio

www.youtube.com/kognitio

info@kognitio.com

About Kognitio

For more than a generation, Kognitio has been a pioneer in software for advanced analytics, helping companies gain greater insight from large and complex volumes of data with low latency and limitless scalability for competitive business advantage. Sitting at the nexus of Big Data, in-memory analytics and cloud computing, Kognitio extends existing data and BI investments as an analytical accelerator, providing a foundation for data scientists and analytical information services. The Kognitio Analytical Platform can be used as a data science lab or to power comprehensive digital marketing analytics; it runs on industry-standard servers, as an appliance, or in Kognitio Cloud, a ready-to-use analytical Platform-as-a-Service (PaaS) in a public or private cloud environment. To learn more, visit kognitio.com and follow us on Facebook, LinkedIn and Twitter.