Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

informatica
Put potential to work.™

Data Warehouse Optimization using Hadoop

A new solution from Capgemini, implemented in partnership with Informatica, Cloudera and Appfluent, optimizes the ratio between the value of data and storage costs, making it easy to take advantage of new big data technologies.

Today's escalating data volumes can prevent Online Transaction Processing (OLTP) systems from generating and processing transactions efficiently which can cause Data Warehouse (DW) performance issues when querying data. Total Cost of Ownership (TCO), too, escalates rapidly because of the need for upgrades to DW hardware and for additional licenses since organizations often pay more to store information simply because they **might** need it.

Radically new capabilities are needed to enable DW and OLTP to function cost-effectively in this new environment. First, organizations need to cope with data volumes that traditional DW platforms (whether RDBMSs[1] or appliances) were never designed for. Second, they must deal with unstructured, semi-structured and structured data.

Big data technologies such as Apache Hadoop excel at managing large volumes of unstructured data. However, this data needs to be pulled together with structured data for analytical purposes. As Big Data technologies and Apache Hadoop are coming into mainstream use, being able to integrate these new technologies with existing legacy Data Warehouse platforms to get the best of both worlds is key.

1 Relational Database Management System

Our solution: seamlessly combine the DW with big data technologies

In partnership with Informatica, Cloudera and Appfluent, Capgemini has developed an integrated solution that allows OLTP systems and DWs to serve their primary functions efficiently and cost-effectively. Data Warehouse Optimization (DWO) using Hadoop incorporates Cloudera's highly available massively-parallel processing Hadoop Platform, with minimal effort to integrate it in the existing landscape. Appfluent Visibility is used to identify what data needs to be archived and ETL/ELT[2] processes should be offloaded. Informatica Information Lifecycle Management (ILM) offloads onto Hadoop the data and processing that is not needed for day-to-day functions from OLTP and DW systems. When required, archived data is made available to the primary systems in various ways. Front-line services can seamlessly retrieve historical information or use Informatica ILM to restore archived data to production applications.

DWO using Hadoop incorporates elements from **Informatica**, **Appfluent** and **Cloudera**; integrated into a single solution delivered by **Capgemini**.

Elements of our solution

DWO using Hadoop consists of elements from Informatica, Appfluent and Cloudera; integrated into a single solution delivered by Capgemini. Cloudera provides a complete, tested and widely deployed open source distribution of Apache Hadoop, making it available for mainstream adoption.[3]

Informatica Big Data Edition helps clients to start using Hadoop quickly to manage and analyze unstructured and semi-structured data. Users are shielded from complexity so the need for Hadoop-specific skills is limited.

Informatica BigData edition to create ETL/ELT (including complex transformation, DQ rules, profiling, parsing and matching) framework and push all heavy lifting ETL/ELT processing to Hadoop environment.

Informatica ILM Archive and Informatica Data Services (IDS) together help to allocate data optimally between the DW and Hadoop, and then make it available to users in a rapid and seamless manner, wherever it is stored.

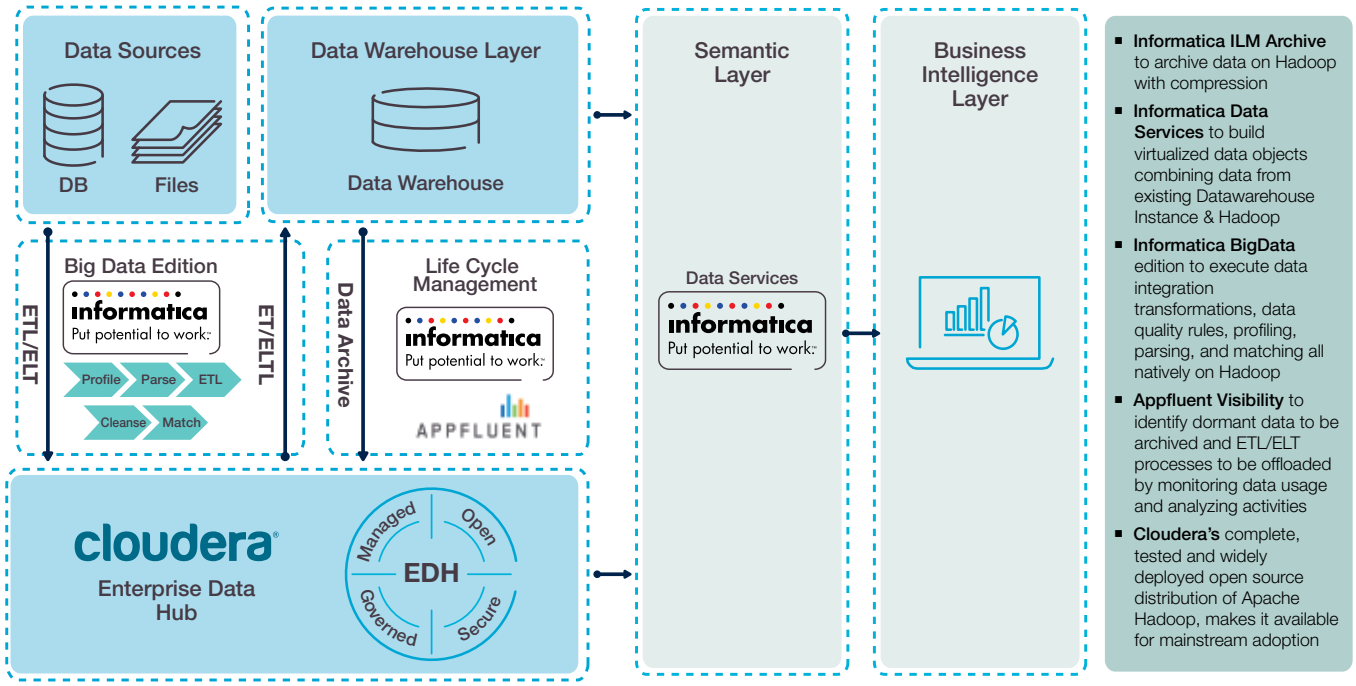
Appfluent Visibility software applies statistical analysis to usage and activity metrics in order to identify dormant data that is a candidate for archiving by Informatica ILM Archive. Hadoop data is held in a highly compressed form, avoiding the need for a regular compression process.

IDS build virtualized data objects that combine data from Hadoop archives and from DWs residing on RDBMSs or appliances. You can get an instant response to your immediate data needs, regardless where the data is stored.

2 ETL: Extract, Transform & Load; ELT: Extract, Load & Transform

3 Please see our brochure Capgemini's Data Optimization for the Enterprise with Cloudera for more details. http://www.capgemini.com/resources/capgeminis-data-optimization-for-the-enterprise-with-cloudera

Figure 1: Seamlessly combine the Data Warehouse with big data technologies



Capgemini expertly delivers DWO using Hadoop – you decide how

As a market leader in big data and business information management, we have the experts to guide you plus the techniques to ensure a successful transformation into a high-performing, information-centric business. We have already helped many clients evolve their legacy information architectures to exploit massive data volumes and new data types.

Below are a few examples of the ways we have helped clients get started.

Strategic Value Assessment (SVA)

In a few weeks, Capgemini can help clients build a phased plan for implementing ILM on Hadoop while identifying key opportunities to optimize the information ecosystem across OLTP, DW and other information assets within the organization. The SVA can also help build a business value proposition and budget to help obtain approval and funding. The phased plan often includes an initial proof of concept (POC) or prototyping phase to help the client understand the approach's effectiveness and test the business case with minimal risk.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

informatica
Put potential to work.™

POC or prototype

We recommend choosing a low-complexity, high-value POC that can be completed quickly. The objective is to create a working model of one representative case. We leverage our relationship with Informatica and Cloudera to conduct POCs at an optimized cost, and can bundle the total cost into one transaction to avoid the complexity of dealing with multiple vendors.

Managed implementation

We can deliver your DWO solution out of our Big Data Service Center: a comprehensive framework leveraging our Rightshore® approach, and bringing together big data strategy governance and organization with an industrialized, high-performance delivery and support capability.

Benefits of DWO using Hadoop

DWO using Hadoop has several key advantages, starting with **improved OLTP and DW performance**. Other benefits include:

TCO reduction: DW upgrade costs are avoided and license costs for existing DWs are reduced because less data needs to be stored there. Commodity hardware and software is used for archived data, lowering infrastructure costs – and archives can be compressed by up to 90%.

Better decision support: All the information you need to make decisions is readily available, together with a full range of analytics. Intelligent archiving combined with virtualization gives optimum performance. Unstructured and structured data can be combined for inclusion in any report or dashboard.

Historical data comes to the forefront of analysis: Respond fast to government and regulatory requests for data.

Increased return on existing investment: You build on the technology you already have, rather than replacing it - your DW becomes future-proof, with infrastructure that scales to support the required volumes. Existing ETL skills can be applied to new technologies.

BI flexibility and stability: A single abstract layer supports any future BI visualization tools and makes it easy to add information in future. There is no change to the business definitions and programming logic of the existing BI structure.

Better data security and governance: Access to data is tightly controlled via rules applied at the semantic layer, and is auditable at a detailed level. Archive files are available only to authorized users. Security and retention policies can be defined across the data.

Find out more

Contact us to learn more about how we can provide Data Warehouse Optimization using Hadoop for your organization to improve OLTP and DW performance.

About Capgemini

Now with 180,000 people in over 40 countries, Capgemini is one of the world's foremost providers of consulting, technology and outsourcing services. The Group reported 2014 global revenues of EUR 10.573 billion.

Together with its clients, Capgemini creates and delivers business, technology and digital solutions that fit their needs, enabling them to achieve innovation and competitiveness. A deeply multicultural organization, Capgemini has developed its own way of working, the Collaborative Business Experience™, and draws on Rightshore®, its worldwide delivery model.

For further information visit
www.capgemini.com/insights-data

or contact us at
insights@capgemini.com

About Informatica

Informatica Corporation (Nasdaq:INFA) is the world's number one independent provider of data integration software. Organizations around the world rely on Informatica to realize their information potential and drive top business imperatives. Informatica Vibe, the industry's first and only embeddable virtual data machine (VDM), powers the unique "Map Once. Deploy Anywhere." capabilities of the Informatica Platform. Worldwide, over 5,000 enterprises depend on Informatica to fully leverage their information assets from devices to mobile to social to big data residing on premise, in the Cloud and across social networks.

For more information, visit
www.informatica.com

M'COS_GI_AH_20150922