Big Data Analytics Platforms analyze from startups to traditional database players

Ionuţ ŢĂRANU
Bucharest University of Economic Studies
ionut.tanaru@gmail.com

Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. With so many emerging trends around big data and analytics, IT organizations need to create conditions that will allow analysts and data scientists to experiment. "You need a way to evaluate, prototype and eventually integrate some of these technologies into the business," says Chris Curran[1]. In this paper we are going to review 10 Top Big Data Analytics Platforms and compare the key-features.

Keywords: *Big data, In-memory, Hadoop, Data analysis*

1 Introduction

The growth of data – both structure and unstructured – will present challenges as well as opportunities for organisations over the next five years.

With growing data volumes, it is essential that real-time information that is of use to the business can be extracted from its IT systems, otherwise the business risks being swamped by a data deluge. Meanwhile, competitors that use data to deliver better insights to decision-makers stand a better chance of thriving through the difficult economy and beyond. To analyze such a large volume of data, big data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high-performance analytics.

Today's advances in analyzing Big Data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook. The business cases for leveraging Big Data are compelling. For instance, Netflix mined its subscriber data to put the essential ingredients together for its recent hit House of Cards, and subscriber data also prompted the company to bring Arrested Development back from the dead.

Another example comes from one of the biggest mobile carriers in the world. France's Orange launched its Data for Development project by releasing subscriber data for customers in the Ivory Coast. The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users. Researchers accessed the data and sent Orange proposals for how the data could serve as the foundation for development projects to improve public health and safety. Proposed projects included one that showed how to improve public safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment.[2] So it seems that data analysis is a do-or-die requirement for today's businesses. We analyze below notable vendor choices, from Hadoop upstarts to traditional database players.

2. Top 10 Big Data Analytics Platforms

2.1. Actian (**Fig.1.**– Actian Analytics Platform).

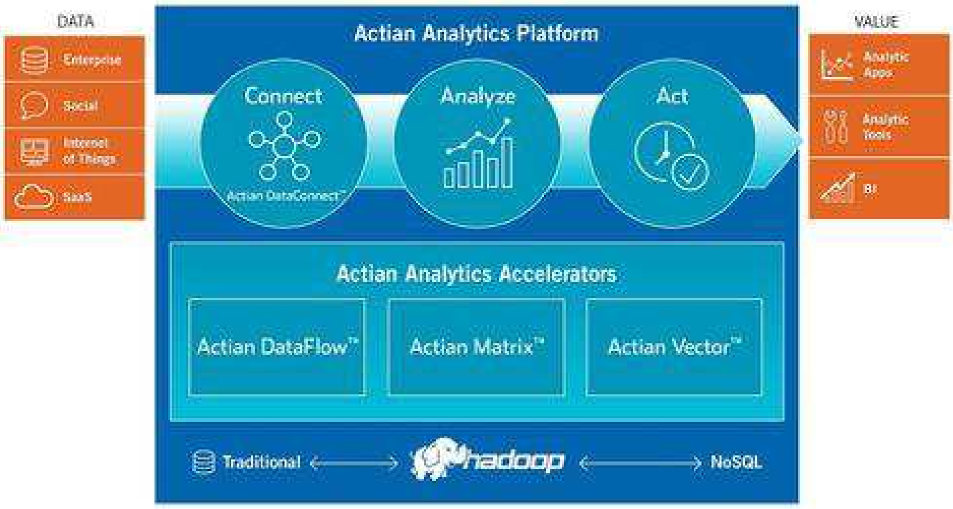

Fig. 1. Actian Analytics Platform

- **Analytical DBMS:** Actian Matrix (formerly ParAccel), Actian Vector (formerly Vectorwise).
- **In-memory DBMS**: Actian Matrix In-Memory Option (data stored to both memory and disk).
- **Hadoop distribution:** None.
- **Stream-processing technology:** None.
- **Hardware/software systems:** None (software-only vendor).

The company is counting on the combination of fast, analytical DBMS options, cloud services, and data-integration and -analytics software geared to a world in which Hadoop is a prominent fixture of the data-management architecture. Actian DataFlow includes SQL-, ETL-, and data-cleansing-on Hadoop options that work with distributions from Apache, Cloudera, Hortonworks, and others [3]

2.2. Amazon

- **Analytical DBMS:** Amazon Redshift service (based on ParAccel engine); Amazon Relational Database Service.
- **In-memory DBMS:** None. Third-party options on AWS include Altibase, SAP Hana, and ScaleOut.
- **Hadoop distributions:** Amazon Elastic MapReduce. Third-party options include Cloudera and MapR.
- **Stream-processing technology:** Amazon Kinesis.
- **Hardware/software systems:** Not applicable.

AWS is located in 11 geographical "regions": US East (Northern Virginia), where the majority of AWS servers are based, US West (northern California), US West (Oregon), Brazil (São Paulo), Europe (Ireland and Germany), Southeast Asia (Singapore), East Asia (Tokyo and Beijing) and Australia (Sydney). There is also a "GovCloud", based in the Northwestern United States, provided for U.S. government customers, complementing existing government agencies already using the US East RegionEach Region is wholly contained within a single country and all of its data and services stay within the designated Region.

Amazon Web Services 2009 (**Fig. 2.**– Amazaon Web Service) hosts a who's who list of data-management services from third-party players -- Cloudera, Microsoft, Oracle, SAP, and many others -- but the cloud giant has its own long-term ambitions where big-data analysis is concerned.[4] Building on its Elastic Compute Cloud (EC2) and Simple Storage Service (S3) storage infrastructure, Amazon launched its Hadoop-based Elastic MapReduce service way back in. In 2013, AWS added the Redshift Data Warehousing service (based on the ParAccel DBMS), which is supported by another who's who list of independent data-integration, business intelligence, and analytics vendors. Rounding out AWS's big-data capabilities are the DynamoDB NoSQL database management service and Kinesis Stream Processing service.

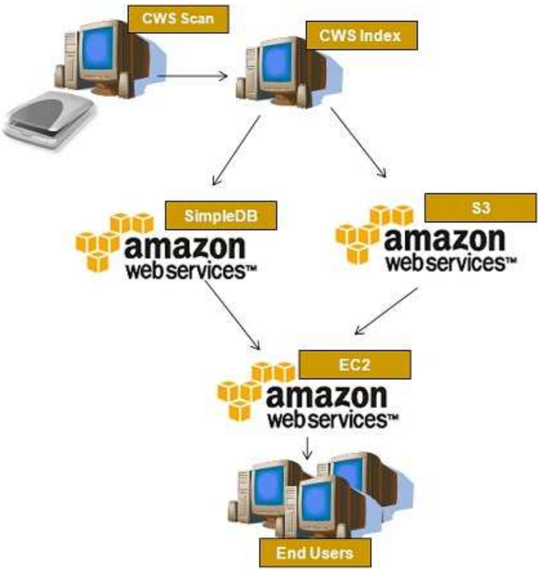

Fig. 2. Amazon Web Service

- Amazon DynamoDB provides a scalable, low-latency NoSQL online Database Service backed by SSDs.
- Amazon ElastiCache provides in-memory caching for web applications. This is Amazon's implementation of Memcached and Redis.
- Amazon Relational Database Service (RDS) provides a scalable database server with MySQL, Oracle, SQL Server, and PostgreSQL support.[22]
- Amazon Redshift provides petabyte-scale data warehousing with column-based storage and multi-node compute.
- Amazon SimpleDB allows developers to run queries on structured data. It operates in concert with EC2 and S3 to provide "the core functionality of a database".
- AWS Data Pipeline provides reliable service for data transfer between different AWS compute and storage services (e.g., Amazon S3, Amazon RDS, Amazon DynamoDB, Amazon EMR). In other words this service is simply a data-driven workload management system, which provides a simple management API for managing and monitoring of data-driven workloads in cloud applications.[23]
- Amazon Kinesis streams data in real time with the ability to process thousands of data streams on a per-second basis. The service, designed for real-time apps, allows developers to pull any amount of data, from any number of sources, scaling up or down as needed.[5]

2.3. Cloudera

- **Analytical DBMS:** HBase, and although not a DBMS, Cloudera Impala supports SQL querying on top of Hadoop.
- **In-memory DBMS:** Although not a DBMS, Apache Spark supports in-memory analysis on top of Hadoop.
- **Hadoop distributions:** CDH open-source distribution, Cloudera Standard, Cloudera Enterprise.
- **Stream-processing technology:** Open-source stream-processing options on Hadoop include Storm.
- **Hardware/software systems:** Partner appliances, preconfigured hardware, or both available from Cisco, Dell, HP, IBM, NetApp, and Oracle.

Cloudera Inc. is an American-based software company that provides Apache Hadoop-based software, support and services, and training to business customers.[6]

Cloudera's open-source Apache Hadoop

distribution, CDH (Cloudera Distribution Including Apache Hadoop), targets enterprise-class deployments of that technology. Cloudera says that more than 50% of its engineering output is donated upstream to the various Apache-licensed open source projects (Apache Hive, Apache Avro, Apache HBase, and so on) that combine to form the Hadoop platform. Cloudera is also a sponsor of the Apache Software Foundation [7]

2.4. HP HAVEn

- **Analytical DBMS:** HP Vertica Analytics Platform Version 7 (Crane release).
- **In-memory DBMS:** Vertica is not an in-memory database, but with high RAM-to-disk ratios the company says it can ensure near-real-time query performance.
- **Hadoop distribution:** None.
- **Stream-processing technology:** None.
- **Hardware/software systems:** HP ConvergedSystem 300 for Vertica, plus a choice of reference architectures for Cloudera, Hortonworks, and MapR Hadoop distributions.

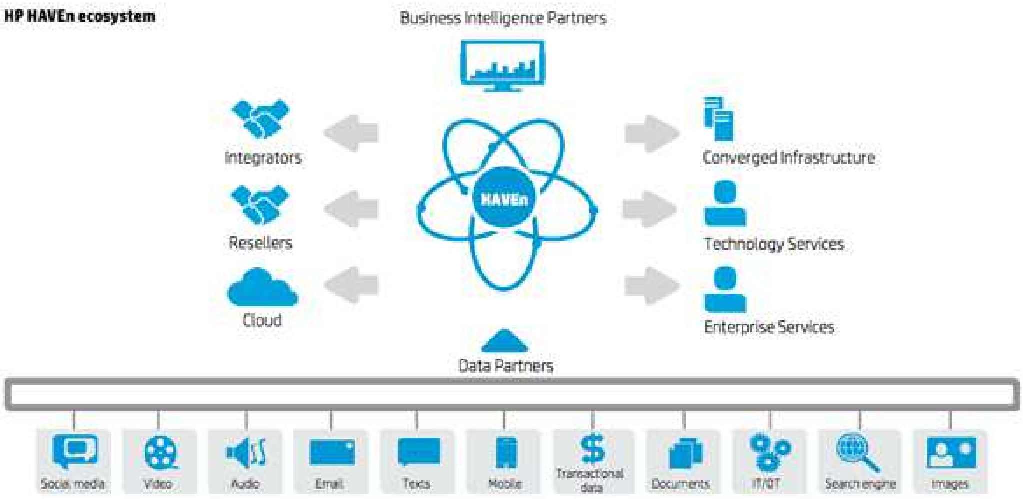

Fig. 3. HAVEn Ecosystem

HP calls its big-data-platform architecture HAVEn (**Fig. 3. -** HAVEn Ecosystem), an acronym for Hadoop, Autonomy, Vertica, Enterprise Security, and "n" applications.

The cluster-based, column-oriented Vertica Analytics Platform is designed to manage large, fast-growing volumes of data and provide very fast query performance when used for data warehouses and other query-intensive applications. The product claims to drastically improve query performance over traditional relational database systems, provide high-availability, and petabyte scalability on commodity enterprise servers.

Its design features include:

- Column-oriented storage organization, which increases performance of sequential record access at the expense of common transactional operations such as single record retrieval, updates, and deletes.[9]
- Standard SQL interface with many analytics capabilities built-in, such as time series gap filing/interpolation, event-based windowing and sessionization, pattern matching, event series joins, statistical computation (e.g., regression analysis), and geospatial analysis.
- Out-of-place updates and hybrid storage organization, which increase the performance of queries,

insertions, and loads, but at the expense of updates and deletes.

- Compression, which reduces storage costs and I/O bandwidth. High compression is possible because columns of homogeneous datatype are stored together and because updates to the main store are batched.[10]
- Shared nothing architecture, which reduces system contention for shared resources and allows gradual degradation of performance in the face of hardware failure.
- Easy to use and maintain through automated data replication, server recovery, query optimization, and storage optimization.
- Support for standard programming interfaces ODBC, JDBC, and ADO.NET.
- High performance and parallel data transfer to statistical tools such as Distributed R, and the ability to store machine learning models, and use them for in-database scoring.[11][12]

2.5. Hortonworks

- **Analytical DBMS:** HBase; although not a DBMS, Hive is Hortonworks' option for SQL querying on top of Hadoop.
- **In-memory DBMS:** Although not a DBMS, Apache Spark supports in-memory analysis on top of Hadoop.
- **Hadoop distributions:** Hortonworks Data Platform (HDP) 2.0, HDP for Windows, Hortonworks Sandbox (free, single-node desktop software offering Hadoop tutorials).
- **Stream-processing technology:** Open-source stream-processing options on Hadoop include Storm.
- **Hardware/software systems:** Partner appliances, preconfigured hardware, or both available from HP, Teradata and others.

On the matter of customer acquisition, six-year-old Cloudera probably has a slight lead over three-year-old Hortonworks (**Fig. 4. -** Hortonworks Data platform), but only just. Analysts estimate Cloudera’s base of paying subscribers at around 350, while Hortonworks’ CEO Rob Bearden says his company has acquired 250 customers over the past five quarters.

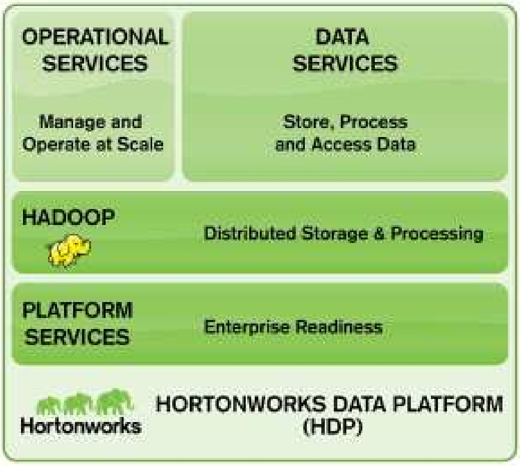
OPERATIONAL SERVICES
Manage and Operate at Scale
DATA SERVICES
Store, Process and Access Data
HADOOP
Distributed Storage & Processing
PLATFORM SERVICES
Enterprise Readiness
Hortonworks
HORTONWORKS DATA PLATFORM (HDP)

Fig. 4. Hortonworks Data platform

The most significant point of disagreement between Cloudera and Hortonworks lies in their answers to a single question – and the one that, arguably, matters most to enterprise customers: should Hadoop complement or replace traditional enterprise data warehouse (EDW) investments?

2.6. IBM

- **Analytical DBMS:** DB2, Netezza (**Fig. 5.** - IBM Netezza platform).
- **In-memory DBMS:** DB2 with BLU Acceleration, solidDB.
- **Hadoop distribution:** InfoSphere BigInsights.
- **Stream-processing technology:** InfoSphere Streams.
- **Hardware/software systems:** PureData System For Operational Analytics (DB2), IBM PureData System for Analytics (Netezza); PureData System for Hadoop (BigInsights).

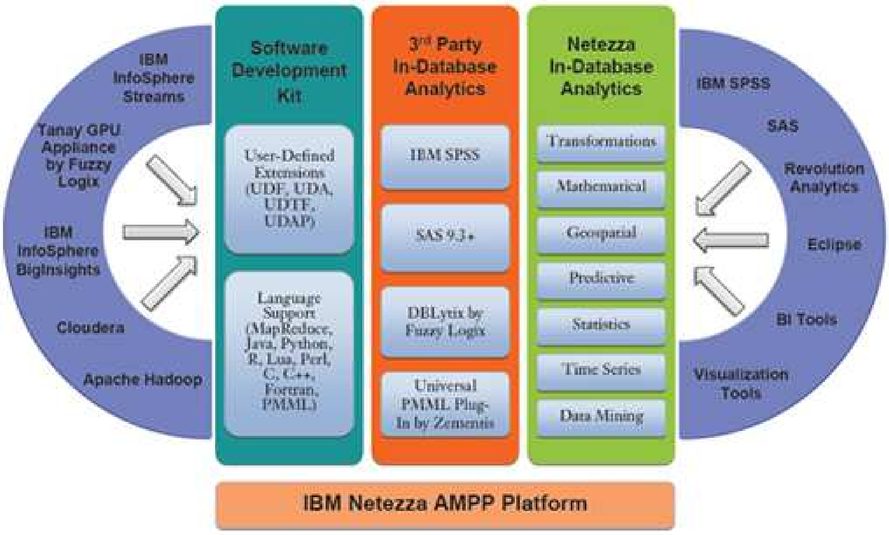
IBM InfoSphere Streams
Tanay GPU Appliance by Fuzzy Logix
IBM InfoSphere BigInsights
Cloudera
Apache Hadoop
Software Development Kit
User-Defined Extensions (UDF, UDA, UDTF, UDAP)
Language Support (MapReduce, Java, Python, R, Lua, Perl, C, C++, Fortran, PMML)
3rd Party In-Database Analytics
IBM SPSS
SAS 9.3+
DBLytix by Fuzzy Logix
Universal PMML Plug-In by Zementis
Netezza In-Database Analytics
Transformations
Mathematical
Geospatial
Predictive
Statistics
Time Series
Data Mining
IBM SPSS
SAS
Revolution Analytics
Eclipse
BI Tools
Visualization Tools
IBM Netezza AMPP Platform

Fig. 5. IBM Netezza platform

Although IBM has plenty of products and services, it's not a product-oriented provider of technology. IBM leads with its deep integration and consulting expertise in a consultative approach focused on building business-differentiating "solutions" that might incorporate multiple products.
IBM Netezza Analytics' advanced technology fuses data warehousing and in-database analytics into a scalable, high-performance, massively parallel advanced analytic platform that is designed to crunch through petascale data volumes. This allows users to ask questions of the data that could not have been contemplated on other architectures. IBM Netezza Analytics is designed to quickly and effectively provide better and faster answers to the most sophisticated business questions. [13]

2.7. Microsoft

- **Analytical DBMS:** SQL Server 2012 Parallel Data Warehouse (PDW).
- **In-memory DBMS:** SQL Server 2014 In-Memory OLTP (option available with SQL Server 2014, set for release by second quarter of 2014).
- **Stream-processing technology:** Microsoft StreamInsight.
- **Hadoop distribution:** HDInsight/Windows Azure HDInsight Service (based on Hortonworks Data Platform).
- **Hardware/software systems:** Dell Parallel Data Warehouse Appliance, HP Enterprise Parallel Data Warehouse Appliance.

The Microsoft Analytics Platform System (**Fig.6.** - Microsoft Analytics Platform System) is a turnkey big data analytics appliance, combining Microsoft's massively parallel processing (MPP) data warehouse technology–the SQL Server Parallel Data Warehouse (PDW)–together with HDInsight, Microsoft's 100% Apache Hadoop distribution, and delivering it as a turnkey appliance. To integrate data from SQL Server PDW with data from Hadoop, APS offers the PolyBase data querying technology.[14]

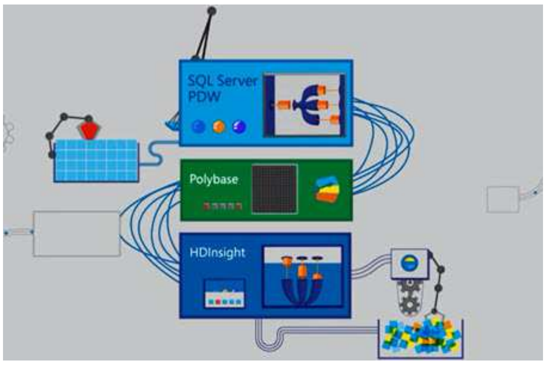

Fig.6. Microsoft Analytics Platform System

2.8. ORACLE

- **Analytical DBMSs:** Oracle Database, Oracle MySQL, Oracle Essbase.
- **In-memory DBMS:** Oracle TimesTen, Oracle Database 12c In-Memory Option (announced in 2013 without details, roadmaps, or release dates).
- **Stream-analysis option:** Oracle Event Processing.
- **Hadoop distribution:** Resells and supports Cloudera Enterprise.
- **Hardware/software systems:** Exadata, Exalytics, Oracle Big Data Appliance.

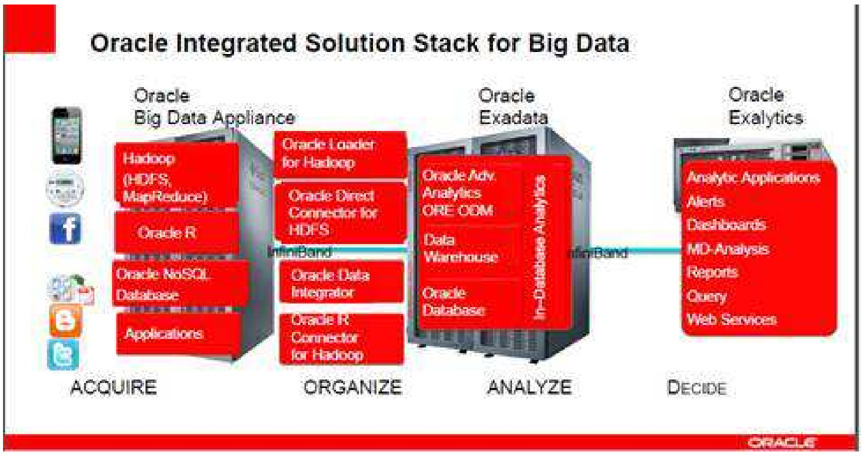

Fig.7. Oracle Big Data Appliance

The **Oracle Big Data Appliance** consists of hardware and software from Oracle Corporation designed to integrate enterprise data, both structured and unstructured. It includes the Oracle Exadata Database Machine and the Oracle Exalytics Business Intelligence Machine, used for obtaining, consolidating and loading unstructured data into Oracle Database 11g. The product also includes an open source distribution of Apache Hadoop, Oracle NoSQL Database, Oracle Data Integrator with Application Adapter for Hadoop, Oracle Loader for Hadoop, an open source distribution of R, Oracle Linux, and Oracle Java Hotspot Virtual Machine [15]

Oracle Big Data Appliance (**Fig.7.** - Oracle Big Data Appliance) By combining the newest technologies from the Hadoop ecosystem and powerful Oracle SQL capabilities together on a single pre-configured platform, Oracle Big Data Appliance is uniquely able to support rapid development of new Big Data applications and tight integration with existing relational data. Oracle Big Data Appliance is pre-configured for secure environments leveraging Apache Sentry, Kerberos, both network encryption and encryption at rest as well as Oracle Audit Vault and Database Firewall.[16]

2.9. Pivotal

- **Analytical DBMS:** Pivotal Greenplum Database.
- **In-memory DBMS:** Pivotal GemFire and SQLFire. Pivotal HD used in combination with GemFire XD and HAWQ for in-memory analysis on top of Hadoop.
- **Stream-analysis option:** Pivotal is working a project aimed at integrating its GemFire (NoSQL) and SQLFire in-memory data grid capabilities with Pivotal Hadoop and Spring XD as a data-ingest mechanism to support scalable, streaming-data analysis.
- **Hadoop distribution:** Pivotal HD.

- **Hardware/software systems:** Pivotal Data Computing Appliance

Pivotal HD is 100% Apache Hadoop compliant and supports all Hadoop Distributed File System (HDFS) file formats. In addition, Pivotal HD supports Apache Hadoop-related projects, including Yarn (aka MapReduce 2.0), Zookeeper and Oozie (for resource and workflow management), Hive and HBase (for language and analytics support).[17]

Pivotal GemFire® stores all operational data compressed and in-memory to avoid disk I/O time lags. Nodes operate in a cluster, optimizing data distribution and processing, to ensure the highest speed and balanced utilization of system resources. Pivotal GemFire scales elastically and linearly – adding nodes increases capacity predictably.[18]

2.10. SAP

- **Analytical DBMSs:** SAP Hana, SAP IQ.
- **In-memory DBMS:** SAP Hana. **Stream-analysis option:** SAP Event Stream Processing.
- **Hadoop distribution:** Resells and supports Hortonworks, Intel; Hadoop integrations certified by Cloudera and MapR.
- **Hardware/software systems:** Multiple hardware configuration partners include Dell, Cisco, Fujitsu, Hitachi, HP, and IBM.

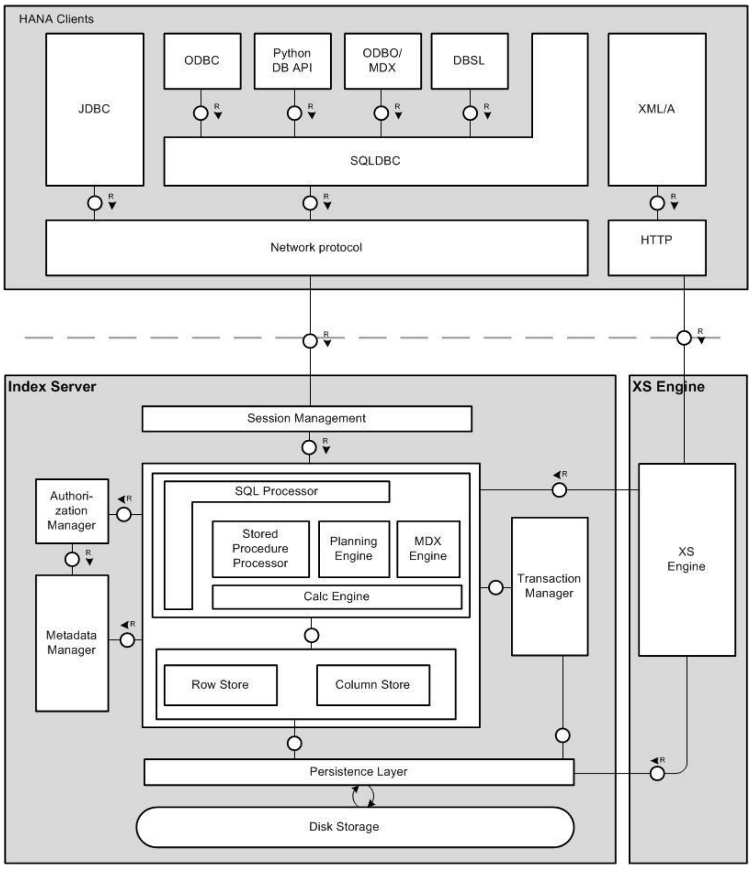

Fig.8 Architecture

SAP HANA is an in-memory, column-oriented, relational database management system developed and marketed by SAP SE. [19] HANA's architecture is designed to handle both high transaction rates and complex query processing on the same platform. SAP HANA was previously called SAP High-Performance Analytic Appliance [20]

The main process, called the index server, has a structure **Fig.8. –** Architecture.

The indexer performs session management, authorization, transaction management and command processing. Note that HANA has both a row store and a column store. Users can create tables using either store, but the column store has more capabilities. The index server also manages persistence between cached memory images of database objects, log files and permanent storage files.

The Authorization manager provides authentication and authorization services. The Authorization Manager can provide security based on SAML, OAuth or Kerberos authentication protocols.

The Extended Services (XS) Engine is a web server with privileged access to the database. Applications written with server-side JavaScript or as Java Servlets can be deployed to the XS Engine. These can either be HTML web applications or REST web service endpoints. Server-side JavaScript includes jQuery-based extensions for database access and to access HTTP request and response messages. The JavaScript engine is based on the Mozilla SpiderMonkey project. [21]

5. Conclusions

With data growing so rapidly and the rise of unstructured data accounting for 90% of the data today, the time has come for enterprises to re-evaluate their approach to data storage, management and analytics. Legacy systems will remain necessary for specific high-value, low-volume workloads, and complement the use of Hadoop -optimizing the data management structure in your organization by putting the right Big Data workloads in the right systems. The cost-effectiveness, scalability, and streamlined architectures of Hadoop will make the technology more and more attractive. In fact, the need for Hadoop is no longer a question. The only question now remaining is how to take advantage of it best. All of these tools provide a rich feature set ready for enterprise use. It will be up to the end user to do a thorough comparison and select either of these tools

References

[1] http://www.networkworld.com/article/2837779/big-data-business-intelligence/8-big-trends-in-big-data-analytics.html.

[2] http://www.datamation.com/applications/big-data-analytics-overview.html

[3] http://www.actian.com/solutions/#customer-analytics-content

[4] http://aws.amazon.com/

[5] http://en.wikipedia.org/wiki/Amazon_Web_Services

[6] http://en.wikipedia.org/wiki/Cloudera

[7] http://www.apache.org/foundation/sponsorship.html

[8] Vance, Ashlee (16 March 2009). "Bottling the Magic Behind Google and Facebook". *The New York Times*.

[9] Monash, C: "Are row-oriented RDBMS obsolete?" *DBMS2*, January 22, 2007

[10] Monash, C: "Mike Stonebraker on database compression – comments”, *DBMS2*, March 24, 2007

[11] Gagliordi, Natalie. "HP adds scale to open-source R in latest big data platform". *ZDNet*.

[12] Prasad, Shreya; Fard, Arash; Gupta, Vishrut; Martinez, Jorge; LeFevre, Jeff; Xu, Vincent; Hsu, Meichun; Roy, Indrajit (2015). "Enabling predictive analytics in Vertica: Fast data transfer, distributed model creation and in-database prediction". *ACM SIGMOD International Conference on Management of Data (SIGMOD)*.

[13] http://www-01.ibm.com/software/data/puredata/analytics/nztechnology/analytics.html

[14] http://www.microsoft.com/en-us/server-cloud/products/analytics-platform-system/

[15] Darrow, Barb (2011-10-03). "Oracle BigData Appliance stakes big claim".

[16] http://www.oracle.com/technetwork/database/bigdata-appliance/overview/index.html

[17] http://pivotal.io/

[18] http://pivotal.io/big-data/pivotal-gemfire

[19] Jeff Kelly (July 12, 2013). "Primer on SAP HANA". *Wikibon*. Retrieved October 9, 2013

[20] http://en.wikipedia.org/wiki/SAP_HANA

[21] https://developer.mozilla.org/en-US/docs/Mozilla/Projects/SpiderMonkey

Mr. Ionuţ Ţăranu graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 1996, having its Master degree on "Database support for business". At present is in the process of getting his title of doctor in economy in the specialty of "Soft-computing methods for early medical diagnosis". He has been an Assistant Professor for 4 years at "Titu Maiorescu" University and also for 4 years at Academy of Economic Studies from Bucharest. He published a series of articles, from which the most important are Applying ABCD Rule of Dermatoscopy using cognitive systems and ABCDE Rule in Dermoscopy – Registration and determining the impact of parameter E for evolution in diagnosing skin cancer using soft computing alghorithms.

Mr. Taranu is currently the General Manager of Stima Soft company. He has more than 15 years of experience as a project manager and a business analyst with over 13 years of expertise in Software development, Business Process Management, Enterprise Architecture design and Outsourcing services. He is also involved in research projects, from which the most relevant are:

- Development of an Intelligent System for predicting, analyzing and monitoring performance indicators of technological and business processes in renewable energy area;
- Development of an eHealth platform for improving quality of life and the personalization of therapy at patients with diabetes;
- Development of an Educational Portal and integrated electronic system of education at the University of Medicine and Pharmacy "Carol Davila" to develop medical performance in dermatological oncology field;