



IBM AND NEXT GENERATION ARCHITECTURE FOR BIG DATA & ANALYTICS

The IBM Big Data Landscape

IBM can legitimately claim to have been involved in Big Data and to have a much broader Big Data portfolio than any other IT vendor, both in respect to software and hardware. It refers to the major part of this swathe of technology as IBM Watson Foundations.

This solution set is so extensive that we make no effort to cover it all in detail here and instead will focus on what we believe to be the core elements. However before we discuss these, it is worth our noting that IBM's approach to Big Data is distinctly different from other major vendors in two respects. The first goes by the name of Watson and the second, Stream Computing.

Watson

Originally, Watson was the computer system IBM built to play against Jeopardy champions in a televised and well-publicized "man vs. machine" contest. That was early 2011, and of course, Watson won. The original project began 5 years earlier, blending together techniques from natural language processing, extensive text search and rules-based artificial intelligence (AI). The resulting technology, which IBM has further developed and enhanced, clearly has application in Big Data. To this end, IBM's Watson Group has married Watson technology with data analytics to create what it refers to as a cognitive computing platform.

Cognitive computing can thus be thought of as the mergence and evolution of several software technologies: natural language processing, knowledge management, AI and data analytics.

Stream Computing and Real-Time Analytics

IBM refers to real-time analytics as stream computing because it requires the processing of data in motion (event streams or data streams). While IBM has a wide variety of technologies to enable and support traditional analytic applications, it clearly recognizes that some analytical activity is extremely time-sensitive, and if the results are to be used effectively it must be actioned immediately or at least very quickly. With real-time analytics of this kind, all the necessary processing of the data needs to be handled while the data is in motion.

Real-time analytics was initially a niche technology taken up in the financial sector because it could be used very effectively in automated trading. It has taken more than a decade for it to move toward the mainstream, but it has begun to do so in the past few years with the increase in publicly available data streams, particularly social media data. Most of the vendors offering technology in this area are niche vendors, but IBM is not. It has integrated its streaming technology into its IBM Watson Foundations.

IBM Watson Foundations

IBM Watson Foundations is an architecture that integrates IBM's portfolio of Big Data and Analytics technology. It can legitimately claim to cover the full spectrum of Big Data activity.

It spans Hadoop technology and all the possibilities of data management, from the traditional relational database and data warehousing to document databases, graph databases and everything else that is normally classified under the NoSQL heading. It even includes data in the form of RDF triples. This is not just about storing data in a variety of data stores, it also involves data governance from data ingest through to data archive, and as we have already noted, it supports the processing of real-time data. It enables the full gamut of business

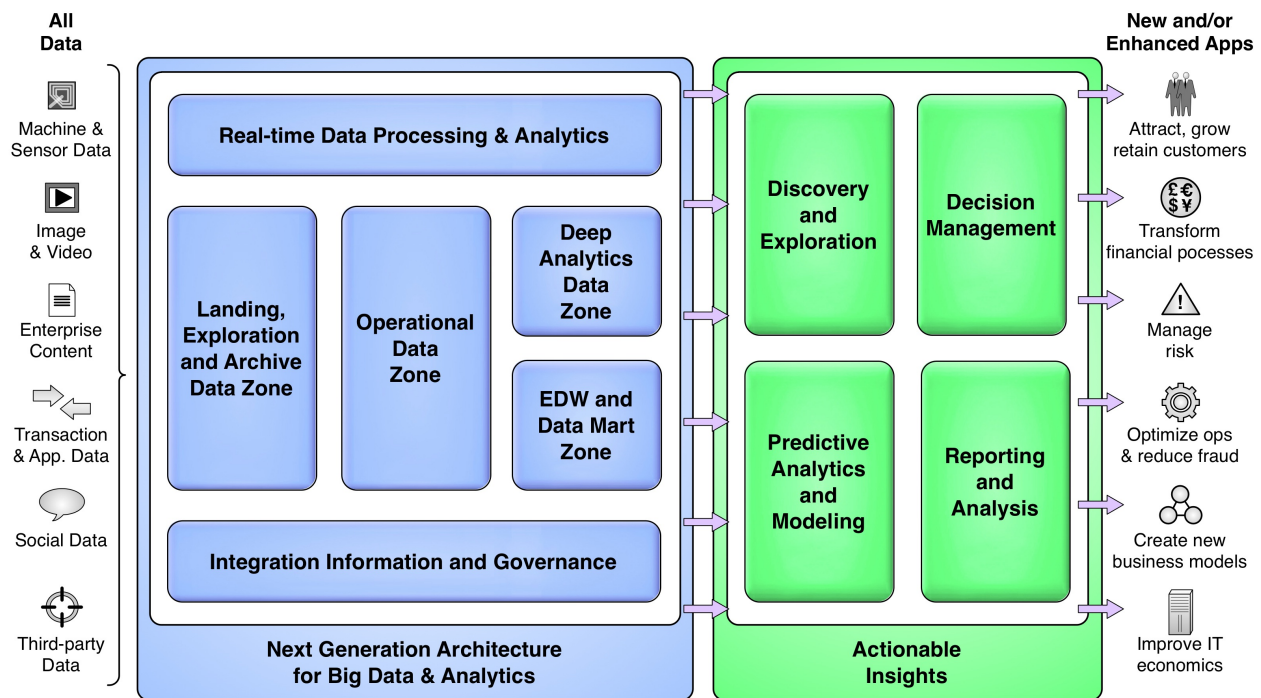


Figure 1: IBM Next Generation Architecture for Big Data & Analytics

intelligence (BI) and analytics applications to which it brings the added power of its Watson technology.

Figure 1 shows IBM Watson Foundations in overview. On the left side of the diagram we see data flowing into the Next Generation Architecture for Big Data & Analytics. Here data is ingested, stored and made available for access. In effect, it constitutes the data layer for Big Data, with various BI and analytics software in Actionable Insights deriving actionable intelligence from it, which in turn can be fed to applications.

Neither the systems of engagement nor systems of record work in isolation, they work together to deliver more value. Combine that with the power of analytics using big data from inside and outside of the organization – enterprise content, social media, enterprise data and machine data – and a solution such as IBM's can transform not only interactions but entire industries.

The various zones of Next Generation Architecture for Big Data & Analytics are:

Integration Information and Governance

This is a broad and unified set of capabilities designed to integrate and govern data. In this area, IBM InfoSphere Server provides massively parallel processing (MPP) to the data integration platform and enables data cleansing, monitoring and transformation. It includes data lifecycle management, which also provides the ability to archive data into Hadoop and perform analytics on multi-structured and archived data. IBM InfoSphere Master Data Management and InfoSphere Entity Analytics Solutions provide management over operational processes and the ability to identify and resolve fraud, and IBM InfoSphere Data

Privacy for Hadoop specifically addresses privacy and compliance initiatives for data stored in Hadoop. It also includes comprehensive data security and monitoring tools via IBM InfoSphere Guardium.

The Landing, Exploration and Archive Data Zone

This is where data lands initially, often directly into Hadoop, where it is examined and possibly analyzed for potential usage. It is also where data that needs to be retained is archived. Here, IBM InfoSphere BigInsights provides an operational environment that hardens Hadoop for enterprise-level deployment. It includes extensive administration, provisioning, security and software development capabilities. It also includes Big SQL, a SQL over Hadoop capability that enables users to explore Hadoop data. As much of the world's big data is unstructured and often in textual content, this area also features text analytics with an impressive library of extractors, a critical component to analyzing and deriving meaning from text.

The Operational Data Zone

This is a data integration and governance area where IBM InfoSphere Information Server provides a comprehensive data integration capability that can flow data directly from Hadoop, or if desired, transfer ETL tasks into a Hadoop infrastructure. InfoSphere Data Click enables users to swiftly load data into and out of Hadoop. InfoSphere Federation Server extends data integration capability with its ability to consolidate virtual views of data across multiple heterogeneous databases (Oracle, SQL Server, etc.).

InfoSphere Information Server also provides a good deal of governance functionality. The Governance Catalog enables users to create, monitor and share business metadata and to search for data through all registered data sources. Components of Cognos BI support the organizational rollout of data governance programs. The Governance Dashboard provides a series of customizable reports for common governance tasks (policies, information governance rules, categories, terms, stewards etc.). InfoSphere Optim provides for data lifecycle management and InfoSphere MDM handles master data management.

InfoSphere Information Server for Data Quality delivers a broad set of data cleansing tools (both batch and real time) and provides a console for data quality monitoring. In addition, IBM Entity Analytics Solutions (EAS) offers a highly sophisticated data wrangling capability for entity detection and analysis, and data disambiguation across people, transactions, relationships and events.

For data security, IBM has InfoSphere Guardium and InfoSphere Data Privacy for Hadoop.

The EDW and Data Mart Zone and the Deep Analytics Data Zone

IBM has an array of database capabilities. DB2, traditionally its data warehouse workhorse, was dramatically enhanced for the Big Data market to become DB2 with BLU Acceleration – a database that is tailored for analytical tasks. It employs a column-store capability, requires minimal DBA overhead and implements many state-of-the-art performance features including extreme compression and on-chip vector instructions.

The IBM Informix database acts as a complementary engine to DB2. Aside from the usual database capabilities, it can store time-series data and documents in JSON-accessible form. As an object-relational database it is highly versatile and can be deployed in multiple roles. IBM also offers IMS on the mainframe, the Netezza Data Warehouse appliances and the FileNet Content manager.

Technically all these database products have been engineered for analytic workloads, and InfoSphere BigInsights includes an impressive amount of analytical functionality covering sophisticated text analytics, social data analytics (for social media data), machine data analytics (for event data) and what IBM calls Big R, its implementation of the R language.

Real-Time Data Processing and Analytics

In this area, IBM InfoSphere Streams delivers an application development environment for building real-time applications that ingest and analyze data streams from multiple sources. It has been built to handle very high data throughput rates to the level of millions of records per second. By analyzing real-time data from sensors, organizations can more effectively anticipate when a problem is likely to occur.

IBM's Actionable Insights

Actionable Insights is where the main BI and analytic software runs that feeds on the Big Data data layer. The four areas described below are in most cases fed data by SPSS data integration capabilities and can leverage IBM's visualization technology, an extensive and free visualization tool that can be viewed and downloaded from AnalyticsZone.com

Discovery and Exploration

This is the area where IBM's Watson technology makes its main contribution to the Big Data picture, via Watson Explorer and Watson Analytics. In combination with other IBM technologies it provides discovery, navigation and search over a wide range of data sources – structured and unstructured – both inside and outside the enterprise. Watson Explorer embodies a framework for developing information-oriented applications that deliver useful, relevant knowledge and information for specific business contexts. Watson Analytics adds in the analytics dimension.

Decision Management

IBM Operational Decision Manager is a platform for capturing, managing and automating repeatable business decisions. It comprises a repository that enables users (subject matter experts) to maintain business decision information and a runtime capability that automates decision logic based on context. This solution is broadened by IBM Decision Optimization Center, which is a configurable platform for feeding relevant analytics to business decision makers. Even greater sophistication in this area is provided by IBM ILOG CPLEX Optimizer, software that allows the modeling of business situations and the applications of algorithms to arrive at precise optimal decisions.

Predictive Analytics and Modeling

IBM's primary software for predictive analytics and a wide variety of other analytical tasks is SPSS. This is available in two principal components: SPSS Modeler, a predictive analytics platform, and SPSS Statistics, which addresses the entire analytics process.

Reporting and Analysis

This area of BI activity is covered by many of the components in the Cognos product set. Aside from regular reporting capability (Cognos BI, Cognos Express), it includes: Cognos TM1, the enterprise planning software platform; Cognos Insight, a personal analytics capability for creating collaborative dashboards and reports; and Concert, a performance management capability. IBM can also provide specific analytics via its Content Analytics and Social Media Analytics products.

The Bottom Line

As we noted initially, IBM has a much broader Big Data portfolio than any other IT vendor, not just in breadth but also in depth – from data acquisition to actionable intelligence. In particular, it has impressive streaming technology and a unique cognitive computing capability.

In our view, any business that is considering an investment in Big Data projects needs to take a look at what IBM can provide.

About The Bloor Group

The Bloor Group is a consulting, research and technology analysis firm that focuses on open research and the use of modern media to gather knowledge and disseminate it to IT users. Visit both www.TheBloorGroup.com and www.OutsideAnalysis.com for more information.

The Bloor Group is the sole copyright holder of this publication.

Austin, TX 78720 | 512-524-3689