ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 3, Issue 12, December 2016

Big Data Architecture: A Survey

[1]Dr. P Bhargavi [2]R.P.L.Durgabai
[1] Assistant Professor [2] Research Scholar
[1][2]Department of Computer Science
Sri PadmavatiMahilaVisvavidyalayam
Tirupati-517 502

***Abstract*:-- Big data architecture describes a dimensions-based approach for assessing the viability of a big data solution. There are 4 key layers of a big data system - i.e. the different stages the data itself has to pass through on its journey from raw statistic or snippet of unstructured data (for example, social media post) to actionable insight. The whole point of a big data strategy is to develop a system which moves data along this path. In this post, I will attempt to define the basic layers you will need to have in place in order to get any big data project off the ground.**

I. INTRODUCTION

A. Data sources layer

This is where the data is arrives at your organization. It includes everything from your sales records, customer database, feedback, social media channels, marketing list, email archives and any data gleaned from monitoring or measuring aspects of your operations. One of the first steps in setting up a data strategy is assessing what you have here, and measuring it against what you need to answer the critical questions you want help with. You might have everything you need already, or you might need to establish new sources.

B. Data storage layer

This is where your Big Data lives, once it is gathered from your sources. As the volume of data generated and stored by companies has started to explode, sophisticated but accessible systems and tools have been developed - A computer with a big hard disk might be all that is needed for smaller data sets, but when you start to deal with storing (and analyzing) truly big data, a more sophisticated, distributed system is called for. As well as a system for storing data that your computer system will understand (the file system) you will need a system for organizing and categorizing it in a way that people will understand - the database. Hadoop has its own, known as HBase, but others including Amazon's DynamoDB, MongoDB and Cassandra (used by Facebook), all based on the NoSQL architecture, are popular too. This is where you might find the Government taking an interest in your activities - depending on the sort of data you are storing, there may well be security and privacy regulations to follow.

C. Data processing/ analysis layer

When you want to use the data you have stored to find out something useful, you will need to process and analyze it. A common method is by using a MapReduce tool (which I also explain in a bit more depth in my article on Hadoop). Essentially, this is used to select the elements of the data that you want to analyze, and putting it into a format from which insights can be gleaned. If you are a large organization which has invested in its own data analytics team, they will form a part of this layer, too. They will employ tools such as Apache PIG or HIVE to query the data, and might use automated pattern recognition tools to determine trends, as well as drawing their conclusions from manual analysis.

D. Data output layer

This is how the insights gleaned through the analysis is passed on to the people who can take action to benefit from them. Clear and concise communication (particularly if your decision-makers don't have a background in statistics) is essential, and this output can take the form of reports, charts, figures and key recommendations. Ultimately, your Big Data system's main task is to show, at this stage of the process, how measurable improvement in at least one KPI that can be achieved by taking action based on the analysis you have carried out. Although people have come up with different names for these layers, as we're charting a brave new world where little is set in stone.

II. LOGICAL LAYERS OF A BIG DATA SOLUTION

Logical layers offer a way to organize your components. The layers simply provide an approach to organizing components that perform specific functions. The layers are merely logical; they do not imply that the functions that support each layer are run on separate machines or separate processes. A big data solution typically comprises these logical layers:

1. Big data sources
2. Data massaging and store layer
3. Analysis layer
4. Consumption layer

Big data sources: Think in terms of all of the data available for analysis, coming in from all channels. Ask the data scientists in your organization to clarify what data is required to perform the kind of analyses you need. The data will vary in format and origin:

- Format— Structured, semi-structured, or unstructured.
- Velocity and volume— The speed that data arrives and the rate at which it's delivered varies according to data source.
- Collection point— Where the data is collected, directly or through data providers, in real time or in batch mode. The data can come from a primary source, such as weather conditions, or it can come from a secondary source, such as a media-sponsored weather channel.
- Location of data source— Data sources can be inside the enterprise or external. Identify the data to which you have limited-access, since access to data affects the scope of data available for analysis.

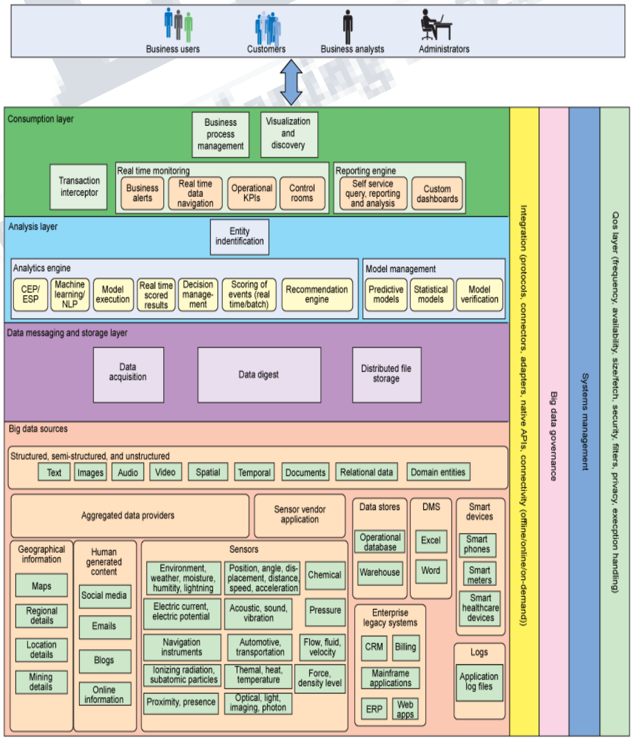
Data massaging and store layer: This layer is responsible for acquiring data from the data sources and, if necessary, converting it to a format that suits how the data is to be analyzed. For example, an image might need to be converted so it can be stored in an Hadoop Distributed File System (HDFS) store or a Relational Database Management System (RDBMS) warehouse for further processing. Compliance regulations and governance policies dictate the appropriate storage for different types of data.

Analysis layer: The analysis layer reads the data digested by the data massaging and store layer. In some cases, the analysis layer accesses the data directly from the data source. Designing the analysis layer requires careful forethought and planning. Decisions must be made with regard to how to manage the tasks to:

- Produce the desired analytics
- Derive insight from the data
- Find the entities required
- Locate the data sources that can provide data for these entities
- Understand what algorithms and tools are required to perform the analytics.

Consumption layer: This layer consumes the output provided by the analysis layer. The consumers can be visualization applications, human beings, business processes, or services. It can be challenging to visualize the outcome of the analysis layer. Sometimes it's helpful to look at what competitors in similar markets are doing. Each layer includes several types of components, as illustrated below.

Figure 1. Components by logical and vertical layer



Big data sources

This layer includes all the data sources necessary to provide the insight required to solve the business problem. The data is structured, semi-structured, and unstructured, and it comes from many sources:

Enterprise legacy systems:— These are the enterprise applications that drive the analytics and insights required for business:

- Customer relationship management systems
- Billing operations
- Mainframe applications
- Enterprise resource planning
- Web applications

Web applications and other data sources augment the enterprise-owned data. Such applications can expose the data using custom protocols and mechanisms.

Data management systems (DMS):— The data management systems store legal data, processes, policies, and various other kinds of documents:

- Microsoft® Excel® spreadsheets
- Microsoft Word documents

These documents can be converted into structured data that can be used for analytics. The document data can be exposed as domain entities or the data massaging and storage layer can transform it into the domain entities.

Data stores:— Data stores include enterprise data warehouses, operational databases, and transactional databases. This data is typically structured and can be consumed directly or transformed easily to suit requirements. Such data may or may not be stored in the distributed file system, depending on the context of the situation.

Smart devices:— Smart devices are capable of capturing, processing, and communicating information on most widely used protocols and formats. Examples include smartphones, meters, and healthcare devices. Such devices can be used to perform various kinds of analysis. For the most part, smart devices do real-time analytics, but the information stemming from smart devices can be analyzed in batch, as well.

Aggregated data providers:— These providers own or acquire the data and expose it in sophisticated formats, at required frequencies, and through specific filters. Huge volumes of data pour in, in a variety of formats, produced at different velocities, and made available by various data providers, sensors, and existing enterprises.

Additional data sources:— A wide range of data comes from automated sources:

1) Geographical information:

- Maps
- Regional details
- Location details
- Mining details

2) Human-generated content:

- Social media
- Email
- Blogs
- Online information

3) Sensor data:

- Environment: Weather, moisture, humidity, lightening
- Electricity: Current, energy potential, etc.
- Navigation instruments
- Ionizing radiation, subatomic particles, etc.
- Proximity, presence, and so on
- Position, angle, displacement, distance, speed, acceleration
- Acoustic, sound vibration, etc.
- Automotive, transportation, etc.
- Thermal, heat, temperature
- Optical, light, imaging, photon
- Chemical
- Pressure
- Flow, fluid, velocity
- Force, density level, etc.
- Other data from sensor vendors

Data massaging and store layer

Because incoming data characteristics can vary, components in the data massaging and store layer must be capable of reading data at various frequencies, formats, sizes, and on various communication channels:

Data acquisition:**—** Acquires data from various data sources and sends the data to the data digest component or stores it in specified locations. This component must be intelligent enough to choose whether and where to store the incoming data. It must be able to determine whether the data should be massaged before it can be stored or if the data can be directly sent to the business analysis layer.

Data digest:— Responsible for massaging the data in the format required to achieve the purpose of the analysis. This component can have simple transformation logic or complex statistical algorithms to convert source data. The analysis engine determines the specific data formats that are required. The major challenge is accommodating unstructured data formats, such as images, audio, video, and other binary formats.

Distributed data storage:— Responsible for storing the data from data sources. Often, multiple data storage options are available in this layer, such as distributed file storage (DFS), cloud, structured data sources, NoSQL, etc.

III. ANALYSIS LAYER

This is the layer where business insight is extracted from the data:

Analysis-layer entity identification— Responsible for identifying and populating the contextual entities. This is a complex task that requires efficient high-performance processes. The data digest component should complement this entity identification component by massaging the data into the required format. Analysis engines will need the contextual entities to perform the analysis.

Analysis engine— Uses other components (specifically, entity identification, model management, and analytic algorithms) to process and perform the analysis. The analysis engine can have various workflows, algorithms, and tools that support parallel processing.

Model management— Responsible for maintaining various statistical models and for verifying and validating these models by continuously training the models to be more accurate. The model management component then promotes these models, which can be used by the entity identification or analysis engine components.

Consumption layer

This layer consumes the business insight derived from the analytics applications. The outcome of the analysis is consumed by various users within the organization and by entities external to the organization, such as customers, vendors, partners, and suppliers. This insight can be used to target customers for product offers. For example, with the business insight gained from analysis, a company can use customer preference data and location awareness to deliver personalized offers to customers as they walk down the aisle or pass by the store.

The insight can also be used to detect fraud by intercepting transactions in real time and correlating them with the view that has been built using the data already stored in the enterprise. A customer can be notified of a possible fraud while the fraudulent transaction is happening, so corrective actions can be taken immediately.

In addition, business processes can be triggered based on the analysis done in the data massaging layer. Automated steps can be launched — for example, the process to create a new order if the customer has accepted an offer can be triggered automatically, or the process to block the use of a credit card can be triggered if a customer has reported fraud.

The output of analysis can also be consumed by a recommendation engine that can match customers with the products they like. The recommendation engine analyzes available information and provides personalized and real-time recommendations.

The consumption layer also provides internal users the ability to understand, find, and navigate federated data within and outside the enterprise. For the internal consumers, the ability to build reports and dashboards for business users enables the stakeholders to make informed decisions and to design appropriate strategies. To improve operational effectiveness, real-time business alerts can be generated from the data and operational key performance indicators can be monitored:

Transaction interceptor— This component intercepts high-volume transactions in real time and converts them into a suitable format that can be readily understood by the analysis layer to do real-time analysis on the incoming data. The transaction interceptor should have the ability to integrate with and handle data from various sources such as sensors, smart meters, microphones, cameras, GPS devices, ATMs, and image scanners. Various types of adapters and APIs can be used to connect to the data sources. Various accelerators, such as real-time optimization and streaming analytics, video analytics, accelerators for banking, insurance, retail, telecom, and public transport, social media analytics, and sentiment analytics are also available to simplify development.

Business process management processes— The insight from the analysis layer can be consumed by Business Process Execution Language (BPEL) processes, APIs, or other business processes to further drive business value by automating the functions for upstream and downstream IT applications, people, and processes.

Real-time monitoring— Real-time alerts can be generated using the data coming out of the analysis layer. The alerts can be sent to interested consumers and devices, such as smartphones and tablets. Key performance indicators can be defined and monitored for operational effectiveness using the data insight generated from the analytics components. Data in real time can be made available to business users from varied sources in the form of dashboards to monitor the health of the system or to measure the effectiveness of a campaign.

Reporting engine— The ability to produce reports similar to traditional business intelligence reports is critical. Ad-hoc reports, scheduled reports, or self-query and analysis can be created by users based on the insight coming out of the analysis layer.

Recommendation engine— Based on the outcome of analysis from the analysis layer, recommendation engines can offer real-time, relevant, and personalized recommendations to shoppers, increasing the conversion rates and the average value of each order in an e-commerce transaction. In real time, the engine processes available information and responds dynamically to each user, based on the users' real-time activities, the information stored within CRM systems for registered customers, and the social profiles for non-registered customers.

Visualization and discovery— Data can be navigated across various federated data sources within and outside the enterprise. The data can vary in content and format, and all of the data (structured, semi-structured, and unstructured) can be combined for visualization and provided to the users. This ability enables organizations to combine their traditional enterprise content (contained in enterprise content managements systems and data warehouses) with new social content (tweets and blog posts, for example) in a single user interface.

Vertical layers

Aspects that affect all of the components of the logical layers (big data sources, data massaging and storage, analysis, and consumption) are covered by the vertical layers:

- ♦ Information integration
- ♦ Big data governance
- ♦ Systems management
- ♦ Quality of service

Information integration

Big data applications acquire data from various data origins, providers, and data sources and are stored in data storage systems such as HDFS, NoSQL, and MongoDB. This vertical layer is used by various components (data acquisition, data digest, model management, and transaction interceptor, for example) and is responsible for connecting to various data sources. Integrating information across data sources with varying characteristics (protocols and connectivity, for example) requires quality connectors and adapters. Accelerators are available to connect to most of the known and widely used sources. These include social media adapters and weather data adapters. This layer can also be used by components to store information in big data stores and to retrieve information from big data stores for processing. Most of the big data stores have services and APIs available to store and retrieve the information.

Big data governance

Data governance is about defining guidelines that help enterprises make the right decisions about the data. Big data governance helps in dealing with the complexities, volume, and variety of data that is within the enterprise or is coming in from external sources. Strong guidelines and processes are required to monitor, structure, store, and secure the data from the time it enters the enterprise, gets processed, stored, analyzed, and purged or archived.

In addition to normal data governance considerations, governance for big data includes additional factors:

- ♦ Managing high volumes of data in variety of formats.
- ♦ Continuously training and managing the statistical models required to pre-process unstructured data and analytics. Keep in mind that this is an important step when dealing with unstructured data.
- ♦ Setting policy and compliance regulations for external data regarding its retention and usage.
- ♦ Defining the data archiving and purging policies.
- ♦ Creating the policy for how data can be replicated across various systems.
- ♦ Setting data encryption policies.

Quality of service layer

This layer is responsible for defining data quality, policies around privacy and security, frequency of data, size per fetch, and data filters:

Data quality

- ♦ Completeness in identifying all of the data elements required
- ♦ Timeliness for providing data at an acceptable level of freshness
- ♦ Accuracy in verifying that the data respects data accuracy rules
- ♦ Adherence to a common language (data elements fulfill the requirements expressed in plain business language)
- ♦ Consistency in verifying that the data from multiple systems respects the data consistency rules
- ♦ Technical conformance in meeting the data specification and information architecture guidelines

Policies around privacy and security

Policies are required to protect sensitive data. Data acquired from external agencies and providers can include sensitive information (such as the contact information of a Facebook user or product pricing information). Data can originate from different regions and countries and must be treated accordingly. Decisions must be made about data masking and the storage of such data. Consider the following data access policies:

- ♦ Data availability
- ♦ Data criticality
- ♦ Data authenticity
- ♦ Data sharing and publishing
- ♦ Data storage and retention, including questions such as can the external data be stored? If so, for how long? What kind of data can be stored?
- ♦ Constraints of data providers (political, technical, regional)
- ♦ Social media terms of use

Data frequency

How frequently is fresh data available? Is it on-demand, continuous, or offline?

Size of fetch

This attribute helps define the size of data that can be fetched and consumed per fetch.

Filters

Standard filters remove unwanted data and noise in the data and leave only the data required for analysis.

IV. SYSTEMS MANAGEMENT

Systems management is critical for big data because it involves many systems across clusters and boundaries of the enterprise. Monitoring the health of the overall big data ecosystem includes:

- ♦ Managing the logs of systems, virtual machines, applications, and other devices
- ♦ Correlating the various logs and helping investigate and monitor the situation
- ♦ Monitoring real-time alerts and notifications
- ♦ Using a real-time dashboard showing various parameters
- ♦ Referring to reports and detailed analysis about the system
- ♦ Setting and abiding by service-level agreements
- ♦ Managing storage and capacity
- ♦ Archiving and managing archive retrieval

- Performing system recovery, cluster management, and network management
- Policy management

Summary

For developers, layers offer a way to categorize the functions that must be performed by a big data solution, and suggest an organization for the code that must address these functions. For business users wanting to derive insight from big data, however, it's often helpful to think in terms of big data requirements and scope. Atomic patterns, which address the mechanisms for accessing, processing, storing, and consuming big data, give business users a way to address requirements and scope. The next article introduces atomic patterns for this purpose.

REFERENCE

1. Min Chen, Shiwen Mao, Yin Zhang, Victor CM Leung , 2014. In Text Book Big Data: Related Technologies, Challenges and Future Prospects

2. Thomas Davenport, 2013. Text Book. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities

3. Stephens ZD, SY Lee, F Faghri, RH Cambell and C Zhai.2015. Text Book. Big data: astronomical or genomical?

4. Tekiner, F.; Keane, J.A., "Big Data Framework," Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on , vol., no., pp.1494, 1499, 13-16 Oct. 2013.

5. VenkatareddyKonasani,MukulBiswas and Praveen Krishnan Koleth, "Fraud Management using Big Data Analytics", A Whitepaper by Trendwise Analytics.

6. Tekiner, F.; Keane, J.A., "Big Data Framework," Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on , vol., no., pp.1494, 1499, 13-16 Oct. 2013.

7. K. Priyanka and KulennavarNagarathna, " A Survey on Big Data Analytics in Health Care", International Journal of Computer Science and Information Technologies, Vol. 5(4),2014, pp. 5685-5688 [6] W. Liu and E.K. Park, " Big Data as an e-Health Service" ,

8. Zaiying Liu; Ping Yang; Lixiao Zhang, "A Sketch of Big Data Technologies," Internet Computing for Engineering and Science (ICICSE), 2013 Seventh International Conference on , vol., no., pp.26,29, 20-22 Sept. 2013

9. Yadav R; Rathod J; V Nair. "Big data meets small sensors in precision agriculture". International Journal of Computer Applications.0975-1,4.2015.

10. VenkatareddyKonasani, MukulBiswas and Praveen Krishnan Koleth, "Fraud Management using Big Data Analytics", A Whitepaper by Trendwise Analytics.

11. SoumendraMohanty, MadhuZaiying Liu; Ping Yang; Lixiao Zhang, "A Sketch of Big Data Technologies," Internet Computing for Engineering and Science (ICICSE), 2013 Seventh International Conference on , vol., no., pp.26,29, 20-22 Sept. 2013.

12. Geraldin B. Dela Cruz, Bobby D. Gerardo, and Bartolome T. Tanguilig III, "Agricultural Crops Classification Models Based on PCA-GA Implementation in Data Mining", International Journal of Modeling and Optimization, Vol. 4, No. 5, October 2014.

13. HemlataChanne, Sukhesh Kothari, DipaliKadam, "Multidisciplinary Model for Smart Agriculture using Internet-of-Things (IoT), Sensors, Cloud-Computing, Mobile-Computing &Big-Data Analysis",Int.J.Computer Technology &Applications,Vol 6 (3),374-382.

14. Kiri L. Wagstaff, Dominic Mazzoni , Stephan R. Sain ,"HARVIST: A System for Agricultural and Weather Studies Using Advanced Statistical Methods",

15. Ch. MallikarjunaRao , Dr. A. AnandaRao , N. Madhusudhana Reddy, "Analysis of Various Crop Yields in Different Spatial Locations of Karimnagar District in AP", IJCSI International Journal of

Computer Science Issues, Vol. 11, Issue 4, No 2, July 2014.

16. SabarinaK ,Priya N, “Lowering Data Dimensionality in Big Data For The Benefit of Precision Agriculture”, Procedia Computer Science 48 (2015) 548 – 554 1877-0509 © 2015 , Published by Elsevier (ICCC 2015) doi: 10.1016/j.procs.2015.04.134.