Building Big Data Analytics Solutions In The Cloud With Tools From IBM

The basic softwares to build own basic big data analytics system on own servers has been discussed on this website through multiple tutorials and we have a list of such big data tutorials for the beginners. Those guides are good to test for the developers including a webmaster on need for log analysis of web server. The running cost of few public servers with 16 GB+ RAM with modern processor is not practical. Additionally there are other type users whose need are not even closest which our tutorials can ever meet. IBM has good resources for our common need (we never used their really complicated service to provide opinion) like guides, projects on Github for some need for setup or development. Cost is not really higher. Apache Hadoop as a service cost $0.70 per hour.

When we are writing about building big data analytics solutions in the cloud, there are few matters we should discuss which are technical or practical in nature for decision making.

The Business Drivers

Often, many companies have at least supportive man power to construct a full fledged big data analysis platform built on the top of commodity servers with attached storage running Hadoop or Cassandra frameworks. For the smaller companies, on-premise hardware is not even really present – they use rented servers, that is the case we provided example with “webmaster” at the beginning of this article. Cloud is also cost-effective compared to on-premises storage for the large business for the additional cost of networking hardware and ISP services. Commonly these larger companies need a hybrid architecture. Part of the total infrastructure has many Free Softwares like Apache Hadoop. Low up-front cost is one of the reason to consider the cost effectiveness.

As the cloud services has ready and running systems, their delivery model enables the clients to rapidly integrate analytics infrastructure with their own (or simply use) without time investment behind provisioning, fine tuning settings, testing etc matters. It is a faster way to use. Of course there are matters like scaling up or down as per need and budget. It is not cost effective to make a 128 RAM dedicated to a smaller cheaper server.

For production environment, there is need of changing analytics capabilities for the new technologies. Cloud providers has service catalogues and usually a dedicated support as a part of managed service.

Similarly, as data volumes grow, new technologies are emerging that deliver scale-out data transfer, enabling efficient, large-scale workflows for ingesting, sharing, collaborating, and exchanging big data.

Security matters within a hybrid cloud or private cloud environment not exactly a concern (we mostly talk about security for the shared public cloud services, mostly freemium in nature). The industries who need security at multiple layers for compliance to industry standards, they can adapt different set of solutions.
Paid services usually have innovative technologies which make the things easier starting from the web GUI.
And lastly, nowadays startups and small businesses are using services which are comparable with the large industries.
There are commonly one or more of the five leading factors which draws a company to plan building big data analytics solutions in the cloud :

1. Big Data Exploration : Finding a data visualizing solution for the need to analyze data for better understanding to improve decision making.
2. 360º View of the Customer : Data analysis can extend the view of the existing customer by incorporating additional internal and external information sources.
3. Security and Intelligence Extension : Some needs to lower a risk, like detecting fraud or monitor cyber security in real time.
4. Operations Analysis : Some of the seekers need is to analyze a variety of data to improve their business results.

5 Modernization of Data Warehouse : Many companies are inclined to integrate enterprise grade big data warehouse capabilities to increase efficiency or enable new types of analysis.

Use Cases

Industry to industry use cases and decisive factors widely vary. From various data sources, Telco seeking the big data analytics solutions mainly for innovative business model, operational efficiency, real time analytics and decision making. Retail are inclined towards personalized recommendations, dynamic pricing and in-store experience. Finance sector including banking seeking solution for real-time fraud detection, alerting, analysis of transaction and behavioral data that is collected, for targeted sales and marketing campaigns and of course for compliance. Both Healthcare and Government has wider usage of Big Data analytics – from analyzing genomics to census. As the volume of data available is growing, the importance of Big Data analytics is increasing. With the new technical methods, the volume of the mostly unstructured data can be transformed into a sensible flow of information. Big data analysis generates an amazing value and influence the structures, processes and management of an organization in a positive sense.
Data now collected from different sources with the aim of generation an economic benefit of the company. Thus points the companies towards the economically meaningful acquisition of Big Data and use of tools for

decision relevant findings from qualitatively diverse and differently structured information, which are subject to rapid change and are incurred to an unprecedented extent. Companies need to understand how to integrate analytics into their everyday operations, structure their organizations to optimize processes, uncover opportunities to stand out from the rest. Here lies the appropriate business application of different analytic techniques.

Due to ever-advancing technological developments and digitalization, companies and organizations have more and more opportunities to collect different data for their benefits through analytical procedures. As a result, the volume of collected data has risen rapidly in recent years. According to a study, approximately 2.5 exabytes of data have been produced since 2012, and this figure is doubling every 40 months. This amount of data was eventually too large to be evaluated by the conventional data base systems. Another problem with the collected data mountains was that they were very different and varied. A uniform data format did not exist, as data were available from more and more sources of different kinds. These different data formats could not be evaluated by a single database kernel, especially the volume of data quickly pushed the data base systems to their limits.

Requirements To Select A Provider

Requirements can be divided in to functional and non-functional. Among the functional requirements, the analytics solution user going to use should support the

wider variant of data sources like CSV, text, XML and so on. The service should have data transformation service to convert data from one format to another format. Provider should have tools to process data sets including capable hardware resources such as suitable storage, compute to stable networking. There should be interactive graphical tools to view the data. It is expected that a provider will include support for data modeling, data warehousing, data repositories, data integration, collections, and of course archiving. The analytics solution must provide end-to-end process, tools, and governance framework for creating, controlling, enhancing, attributing, defining, and managing a metadata schema, model, or other structured aggregation system. There must be metadata management, a centralized repository to create, manage, validate, administer, and monitor analytic models. Of course there is matters like data security, data protection, support of all relevant regulatory matters. It is expected that an analytics solution will have web-based, self-service analytics tools with defined capabilities like for data exploration, discovery to ad-hoc BI queries.
Among the non-functional requirements, there are obvious matters expected from any paid service including a good performance to reliability, availability to fault-tolerance. Now let us assume that a company wants to introduce a BI solution. Here are some points :

1. Flexibility : The solution is intended to achieve independence through the flexibility, which means that the focus of the users is not only placed on the

data access, but the continuously changing requirements and functions can always be made available and implemented as quickly as possible. In the future, demand for complex reporting can grow. The reporting and analysis software should therefore cover not only current requirements, but also potential future needs.

2 Security : One aspect is covered by SSL encryption, and another is question whether the software can allow user and role based security concepts.
3 Learnability : How quickly can a developer learn how to build a custom solution matters.
4 Mobility : Whether the application can be used on various devices including mobile devices. Again, there are different requirements on the part of the end devices.
5 Development capability : The growth of mobile end-users in the business sector is of paramount importance, Such as the trend towards BYOD.
6 Flexibility, security, learnability, mobility, viability are not an exhaustive criterion for assessing reporting and analysis software but a future-oriented software selection.

Architecture

Now our backend in relation to cloud can be from a wide array of choices to deploy and integrate :

1 Public cloud : Easiest and cost effective. Managed Hadoop as a Service is a good example of such

solution like we mentioned at the beginning of this guide.

2. Private cloud : Demands some in-house processing components. A dedicated server or colocation may suffice.
3. Hybrid architecture : Needs enterprise grade on-premise hardware setup and also processing components and work on network part.

Without going in to too much technical details, an analytics solution can be :

1. Basic data platform infrastructure service, such as Hadoop as a Service.
2. Data management services including as a data lake service.
3. An insight and data service

At What Level Of Architecture One Needs To Integrate

As example, for handling threats against a public server, steps in a typical intelligence solution are collecting data from both internal sources such as network probes, DNS, NetFlow, AD logs, and network logs and by external sources such as blacklist and whitelist providers. Most of the structured sources of data are sent to a data integration layer which converts all the incoming data into a single format. Now the next stream computing layer picks up both the streaming flow data such as DNS and NetFlow as well as processed data from the other system. Then the platform computes

simple analytics such as traffic analysis, number of requests made/failed to a DNS, domain etc which are used in developing machine-learning models. All raw data and output are sent to the data repository stored for Apache Hadoop and Machine-learning models are run against longer data sets to detect advanced persistent threats. Additional models from stats language R are also deployed. Machine-learning models that have been developed are deployed which scores them in real time to analyze network, user, and traffic behavior. Custom blacklists from the client and other data sources are used to enrich and pinpoint user activity. Security analysts use custom software interface supplied for Apache Hadoop for visualizations. User look-up information is analyzed to establish exactly which user was involved in a particular traffic flow.

Watson Analytics guides data exploration, automates predictive analytics, and has API support. The common softwares we use by installing on servers to test like Apache Spark, Apache Hadoop, Elasticsearch – they are available as a service (which we have guides for installation on own server). For complicated need it more practical to check their service catalogue. In this article we pointed our regular readers to the usable services for building big data analytics solutions in the cloud for common need in cost effective manner.