International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Rise of AI for IT Operations in Data Lake

Sourabh S Dandagi[1], Girish Rao Salanke N S[2]

[1]Department of Computer Science and Engineering, R.V. College of Engineering, Karnataka, India
[2]Assistant Professor, Department of Computer Science and Engineering, R.V. College of Engineering, Karnataka, India

---

Abstract - *The enterprise infrastructure nowadays has a very complex structure of different physical, virtual and cloud elements. One of them is the Data warehouse. Although data warehouses can hold a large amount of data, it cannot accommodate the entire company's data. In Data lake, all the data of any enterprise is stored at a single location. This new concept of Data lake, will make data operations even easier. The data operations could be pertaining to ingestion of data, removal of data, etc. To manage and monitor these data operations, we have IT Operations. When these IT operations are clubbed with artificial intelligence and machine learning, it automates the task of the IT operations. This could even help in predicting huge road blocks for any applications that's running in the enterprise. By levering this technique, enterprises could automate and enhance their business services. Amalgamating data lake and AIOps (AI for IT Operations), the enterprise will have all the data in a place and this data monitored by AI based IT operations.*

***Key Words*: Data warehouse, Data lake, AIOps, ITOA.**

1. INTRODUCTION

The data warehouses are a well-designed system. The data model of the warehouse is carefully designed according to the usage of the application that has to be deployed. After the warehouse is up and running, the data is loaded. Thousands of users can access the data in the warehouse simultaneously for performing various tasks like reporting or analytics. The data present in the warehouse supports only batch workloads.

Data warehouses were created to combine information from various different data sources so that evaluation and analytics of the application could be visible for everyone. This would result in a single form of the application. Any changes or improvisations for the application could be done by the accessible end users and it would reflect everywhere. This was one of the major advantages of data warehouse.

Though data warehouse concept was a huge success for siloed data, it had several road blocks. The inter-communication between two data warehouses could cause overload on the performance. Enterprises target in meeting the SLA's and a faster response of the application is common in most of them. So, when the application wants to access data from a different warehouse or any data that doesn't fall into the same warehouse, the warehouse has to communicate externally. The response time for this could vary based on the condition of the external devices.

The IT infrastructure of any enterprise has different kinds of data flowing into its system. In data warehouses, the data that is going to flow will be specified and only that particular data flows through. So, when the data warehouse encounters a data which is not compatible with its environment, it has to be converted and then utilized for the required jobs. This could also result in an overhead for the warehouse, causing a delay in response time.

The data today is divided into structured, semi-structured, and unstructured data. Accommodating all these different data into one single warehouse define the Data Lake. It is a place to accumulate every type of data in its instinctive format with no constraints on account size or file. It offers huge data quantity to increase analytic performance and innate integration. Data Lake is similar to a large container which is very similar to real lake and rivers. Just alike a lake, you have multiple streams coming in, a data lake has structured data, unstructured data, machine to machine, logs moving through in real-time. The Data Lake designates data and is a cost-effective way to store all data of an organization for post processing.[1]. Research Analyst can concentrate on finding meaning patterns in data and not data itself. Unlike a hierarchal Data warehouse where data is stored in Files and Folder, Data lake is having a flat architecture. Every element in a Data Lake is given a unique identifier and is tagged with a set of metadata information.

The enterprise looks for a smooth functioning of their application from the consumer's end. Since all the data is consolidated at a place, the applications running in the environment can access any data and the response time would also reduce.

Enterprises have noticed these capabilities of Data Lake and are investing in making a data lake for themselves. This could help in giving a greater visibility, network and operation infrastructure of the whole system.

The data lake should be managed and monitored to prevent it from any anomalies or any deadlock that could shut down the application.[2]. The IT operations look into this monitoring and managing the data facilities. The method that most of the companies follow is IT Operations Analytics (ITOA). ITOA involves analyzing and monitoring of data that is already in the data lake. It diagnoses the data when after the data is received and gives suggestions for the improvement.

ITOA is designed for only some specific kind of data. For example, when there are spark jobs running and there is ITOA tool to look into these jobs and provide the monitoring services, this same tool cannot be used for the data that is coming from Hive or Kafka. AIOps on the other hand has an upper edge over such conditions.

AIOps expands on ITOA in three primary ways:
1) Ingesting more kinds of data
2) Processing real-time data as well as historical data
3) Introducing machine learning to help analyse growing data sets.

2. DATA WAREHOUSE AND DATA LAKE

The Table 1 shows the comparison of data warehouse and data lake.[3]. The key points that come in choosing one technology are structure, process, users, and overall agility of the model to be unique. When these requirements are known, then selecting of the technology would be more profitable for the enterprise.

Gartner had stated that by 2023, 30% of the enterprises would opt for data lakes. This shows how data lake is emerging in the industry due to its advantages.[4]

3. IT OPERATIONS IN ENTERPRISES

There are four basic stages that are involved in IT operations monitoring. They are shown in the Figure 1. Depending on the enterprise capability of monitoring its operations, it follows different stages.

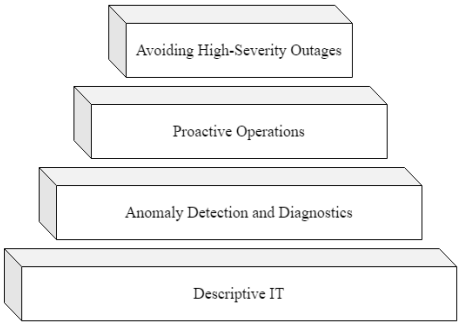

Fig -1: Stages of IT Operations

- Descriptive IT: The data that is pulled into the Data Lake is taken into consideration for analysis. The statistics that is prepared by the existing data helps in monitoring any application.
- Anomaly detection and diagnostics: If any process shows a pattern that leads to an anomaly, the administrator takes care of that before it reaches an unstoppable stage.

Criteria	Data Warehouse	Data Lake
Type of Work Loads	Concurrent users (approx. 1000's). Interactive Analytics Load Management Capabilities. Batch processing	Batch processing. Improvising the capabilities based of users based on the data present
Schema	Schema is well defined before the data is stored. It's called Schema on write; where the data is identified and the model is made. Offers performance, security and integration of the data that it holds. When the type of data is known then the work has to be put in advance.	Schema is defined after the data is entered. It's called Schema on read; the data has to be identified by the code to access the data. There is extreme agility and ease of data capture. The work is required at the end of the data storing. There is no restriction for the use of any particular data type.
Scale	At moderate cost, it can scale large data volumes	At very low cost, it can scale extremely large volume
Data Accessing	Supports SQL and standard BI tool, which are used for reporting and analytics	Programs created by the architects to access data, similar to SQL systems.
Query	Standard SQL	Based programming done by the developer
Advantages	Fast response time within the same warehouse. Consistent performance. Ease of accessing the data. Secure data. Build once, use many	Unmatchable scalability. Parallelization of programming languages Supports Pig and HiveQL. Can store very huge volumes of data
Data	Cleansed data	Raw data
Access	Seeks	Scans
Complexity	Complex joins	Complex processing
Cost/Efficiency	Efficient use of CPU/IO Higher cost	Low cost storage and processing

Table -1: Comparison of Data Warehouse and Data Lake

- Proactive operations: The tool itself is given the privilege to identify any anomaly patterns in the environment and declare what action has to be taken against it.
- Avoiding high-severity outages: Certain anomalies cannot be sited by an administrator. Training the tool in such a way that will help us identify such anomalies will help the applications run smoothly.

4. IT OPERATIONS ANALYTICS (ITOA)

IT operations analytics (ITOA) is like AIOps platforms. Data is aggregated by ITOA tools from various sources and big data analytics is applied to extract the insights. This sort of analytics that was present before AIOps. This kind of analytics was used by the IT operations teams in various organizations, Gartner describes several applications of ITOA systems:

- Root Cause Analysis: All the models, structures and pattern of the IT infrastructure or the specific application can be monitored which can help the users figure out the root causes of the overall system behaviour pathologies.
- Proactive Control of Service Performance and Availability: Predicts the impact of the states, which were formed by the system.
- Problem Assignment: Gives a way of determining the solution or, gives inferences to the most appropriate person or the team in the company to resolve the problem.
- Service Impact Analysis: The relative impact of any of the root causes that has already taken place is figured out. This would help in devoting the resources to correct the fault in least time and cost effectively.
- Complement Best-of-breed Technology: All the models, structures and pattern of the IT infrastructure or the specific application can be monitored are used correct or extend the results of the various discovery-oriented tools to improvise fidelity of information used in the tasks. (e.g., service dependency maps, application runtime topologies, network topologies).
- Real time application behaviour learning: Based on user pattern and the company's infrastructure, it tries to learn and correlate the behaviour of the applications like application patterns, create metrics of such correlated patterns and use it for future analysis.
- Dynamically Baselines Threshold: Based on user pattern and the company's infrastructure, it tries to learn and correlate the behaviour of the infrastructure on various application user patterns and determines the Optimal behaviour of the infrastructure and technological components and changes them accordingly without manual intervention.[5]

5. AIOPS INSIGHTS

5.1 Key Components of AIOps

- Monitoring Ecosystem: The physical and virtual storage of any enterprise should have a greater extent of visibility and accessibility. This is provided by AIOps. Monitoring these tools are crucial for smooth functioning, high service quality and performance.
- Engagement System: While monitoring the data, there can be a creation of extraneous noise. This component decreases the noise and gives service level insights about this to the assigned people in real-time. The operations team should look into such breaks happen. The problem is detected well before by applying machine learning and pre-emptive measures are taken.
- System of Record: Managing trouble tickets, service requests and stores this information for future references. This also can be improvised when unknown relations are found out.
- System of Automation: The resolution scripts are automatically pre-empted.
- Data Lake: It is a repository for all the diagnostics, ad-hoc reporting and business dashboards.

5.2 AIOps Platform Benefits

- Gives a deeper visibility into the business, IT, network and operations infrastructure.
- The analysis and diagnostics of the issues is all automated.
- For the performance check of the infrastructure, alerts and notifications are provided.
- Automated behaviour prediction is done while the monitoring the data.
- Recommendations are provided by looking into the real-time and historical data.

5.3 Next Generation Solution for IT Operations

The complexity in the databases of every enterprise will keep on increasing day after day and the human capability to identify the flaws will remain the same. Therefore, the IT operations have to leverage AIOps platform to monitor and manage their data. The aim of the AIOps vendors, like Datadog, Splunk, Dynatrace, AccelData, etc, should address the IT complexity and high performance demands of the enterprises. These vendors help in decreasing the response time and efficient use of infrastructure resources.[6][7][8][9].

The proactive monitoring helps to identify and act upon it immediately. The demand for this type of monitoring is increasing. A survey states that 74% of the IT professionals want proactive monitoring and analytics, but 42% of them are still using monitoring tools reactively to find and resolve technical issues. [10]

6. ENTERPRISE AIOPS PLATFORM

The enterprise architecture of the data lake which implements AIOps can be as shown in the proposed Figure 2.

The different raw data that are coming through network agents, API's, streaming and ETL are stored in the data lake. This together we can term as a data collection. This data collection is accessible by the applications running in the data lake. On this data, historical and real-time processing are done. This historic and real-time data is accessed by the AIOps tool for pattern discovery, anomaly detection or for some causality. An UI can be made such a way that the presentation of data is interpretable for the user. This UI can be alienated as per the business applications, DevOps and Operations of the underlying data lake.

Privacy plays a major factor for any enterprise. The only accessible way is through the applications of the enterprise and hence these applications should have the limited access. This layer authenticates the user and allows the usage of the applications. The applications consume as well as provide data to the data lake. The data ingestion through the application should be managed and then taken as an input to the data lake.

This is a proposed architecture of an enterprise level data lake with AIOps for managing and monitoring the data in it.

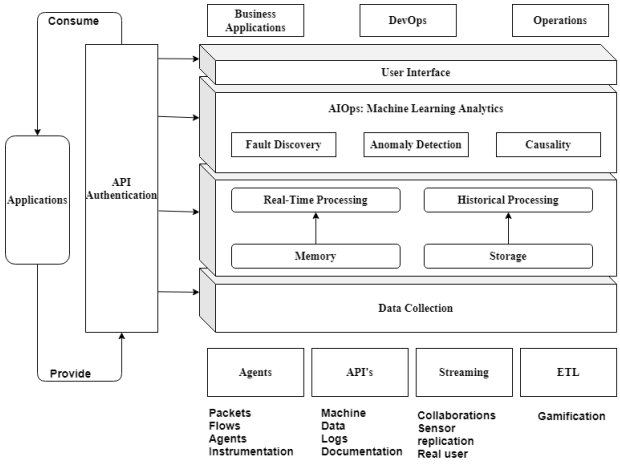

Fig -2: Data Lake Architecture for an Enterprise

7. CONCLUSION

The reactive performance monitoring will create stressful war room situations and damage the company's brand. If there is a road block to run any application due to lack of resources or some other root cause, the enterprise has to engage other developers from different teams to analyse and fix the problems. This would halt the developments of the company.

Adoption of AIOps into data lake of the enterprise will help in various ways like, intelligent alerting to indicate an emerging issue, automated root cause analysis and business impact assessment, and automated remediation for common issues. IT leaders have given an opinion on having proactive approach towards monitoring of the applications. AIOps will prevent any kind of failure of applications running in the data lake. This proactive method of tackling the anomalies will help in reducing the response time thus making the user experience more pleasant.

REFERENCES

[1] Dina Kholar, "The Growing Importance of Data Democratization", 27 Dec 2019, [online], Available: https://www.cfo.com/technology/2019/12/the-growing-importance-of-data-democratization/

[2] Y. Dang, Q. Lin and P. Huang, "AIOps: Real-World Challenges and Research Innovations," *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, Montreal, QC, Canada, 2019, pp. 4-5.

[3] Shana Pearlman, "Data Lake vs Data Warehouse", 10 Oct 2019, [online], Available: https://www.talend.com/resources/data-lake-vs-data-warehouse/

[4] Susan Moore, "How to get started with AIOps", 26 Mar 2019, [online], Available: https://www.gartner.com/smarterwithgartner/how-to-get-started-with-aiops/

[5] "IT Operations Analytics (ITOA) – Optimize IT Operations with Big Data Analytics", 2017, [online], white paper by mayato GmbH, Available: https://www.mayato.com/wp-content/uploads/2017/04/WP-White-Paper-IT-Operations-Analytics-ITOA_EN.pdf

[6] A. Lerner, "*AIOps Platforms*", 09 Aug 2017, [online] Available: https://blogs.gartner.com/andrew-lerner/2017/08/09/aiops-platforms/.

[7] Peter Putz, "AIOps done right: introducing the next generation of software intelligence", 29 Jan 2019, [online], Available: https://www.dynatrace.com/news/blog/aiops-done-right-introducing-the-next-generation-of-software-intelligence/?_ga=2.119102028.435913248.1586817477-1391295656.1586817477

[8] "Everything you need to know about AIOps," Feb. 2019, [online] Available: https://www.moogsoft.com/resources/aiops/guide/everything-aiops/.

[9] Rick Fitz, "AIOps, Machine Learning and the New Era of IT", 18 Aug 2019, [online], Available: https://www.splunk.com/en_us/blog/it/aiops-machine-learning-and-the-new-era-of-it.html

[10] Sonja Jacob, "The Rise of AIOps: How Data, Machine Learning, and AI Will Transform Performance Monitoring", 17 Dec 2018, [online], Available:

https://www.appdynamics.com/blog/news/aiops-platforms-transform-performance-monitoring/

[11] H. Lin et al., "Cloud BOSS: Cloud-centric BSS/OSS for enterprise cloud service operations," 2011 13th Asia-Pacific Network Operations and Management Symposium, Taipei, 2011, pp. 1-4.

[12] D. E. O'Leary, "Embedding AI and Crowdsourcing in the Big Data Lake," in IEEE Intelligent Systems, vol. 29, no. 5, pp. 70-73, Sept.-Oct. 2014. doi: 10.1109/MIS.2014.82

[13] H. Fang, "Managing data lakes in big data era: What's a data lake and why has it become popular in data management ecosystem," 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), Shenyang, 2015, pp. 820-824.

[14] Rick Fitz, "What Is AIOps and What It Means for You", 16 Nov 2017, [online], Available: https://www.splunk.com/en_us/blog/it/what-is-aiops-and-what-it-means-for-you.html

[15] Fernando Iafrate, "Artificial Intelligence and Big Data: The Birth of New Intelligence, Volume 8", Wiley, 2018.

[16] R. Raju, R. Mital and D. Finkelsztein, "Data Lake Architecture for Air Traffic Management," 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, 2018, pp. 1-6.

[17] Y. Chen, H. Chen and P. Huang, "Enhancing the data privacy for public data lakes," 2018 IEEE International Conference on Applied System Invention (ICASI), Chiba, 2018, pp. 1065-1068.

[18] Masood, Adnan & Hashmi, Adnan. (2019). "AIOps: Predictive Analytics & Machine Learning in Operations." 10.1007/978-1-4842-4106-6_7.

[19] S. Nedelkoski, J. Cardoso and O. Kao, "Anomaly Detection and Classification using Distributed Tracing and Deep Learning," 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Larnaca, Cyprus, 2019, pp. 241-250.

[20] U. Thakore, H. V. Ramasamy and W. H. Sanders, "Coordinated Analysis of Heterogeneous Monitor Data in Enterprise Clouds for Incident Response," 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Berlin, Germany, 2019, pp. 53-58.

[21] A. Levin *et al.*, "AIOps for a Cloud Object Storage Service," *2019 IEEE International Congress on Big Data (BigDataCongress)*, Milan, Italy, 2019, pp. 165-169.

[22] Rick Fitz, "3 Ways to Get Started With an AIOps Solution", 09 Mar 2019, [online], Available: https://www.splunk.com/en_us/blog/it/3-ways-to-get-started-with-an-aiops-solution.html

[23] Ryan Covell, "Will AIOps help you solve problems faster?", 25 June 2019, [online], Available: https://www.dynatrace.com/news/blog/will-aiops-help-you-find-the-root-cause/?_ga=2.37860710.435913248.1586817477-1391295656.1586817477

[24] "Part 1: Alerting for the Open-Core Enterprise Data Stack", 23 Aug 2019, [online], Available: https://acceldata.io/blog/f/part-1-alerting-platform-for-the-open-core-enterprise-data-stack

[25] Vikram, "AIOps: Forecasting Cloud Capacity", 1 Sept 2019, [online] Available:https://acceldata.io/blog/f/aiops-forecasting-cloud-capacity-requirements