Bansilal Ramnath Agarwal Charitable Trust's
Vishwakarma Institute of Technology
(An Autonomous Institute affiliated to Savitribai Phule Pune University)

Department of Information Technology

IT-Bulletin

Dec-2022

"Everything is going to be connected to cloud and data"

"Azure Arc is the glue that brings the power of Azure – time to market, innovation, security-to all cloud environments"

'Azure Databricks'

HIGHLIGHTS

"Data is the new science; Big Data holds your answers"

Hi ,

I am Zoya Jamadar and I, like exploring cutting edge technologies that deliver state-of-the-art solutions to the world. We are in a phase where AI is a major turning point in every field that will change the way we all work. Conventional thinking and traditional operating models will not be able to capture the opportunities that arise from this paradigm shift. We therefore need to keep up with technology so that even a modest idea can be scaled in the most practical and efficient way. Consider reading this month's article on one of the hottest topics: Azure Databricks. Please let me know if this inspires you or whether you find any point this article may improve.

Zoya Jamadar

TY(AIDS-B)

What's Azure Databricks

In a nutshell Azure Databricks is a highly optimized version of Apache spark, a service that provides Big Data Analytics integrated in Azure and supported by Microsoft that performs wonders for Data Science and Data engineering needs. It's one of the most sought after tools for scaling machine learning and data science models for real world analytic scenarios.

"Microsoft"

"Data is the new science; Big Data holds your answers"

Majority of our projects are well trained neat models that are running somewhere on cloud Notebooks. Now it's worth considering how these might be scaled for a real time production use. Today here let's look at the fundamentals of Apache spark, data bricks, and some technical design concepts that will surely add value to your cloud experience. Let's get started, happy reading!! …

Getting the terms right

Understanding Apache Spark, Databricks and Azure Databricks

Data these days are either stored in cloud or cloud-like private environments, this data volume is growing exponentially due to improved methods of data collection in various forms including information from our mobile phones or medical reports residing in there. It is a tough job to process data at a massive scale, that's when the notion of distributed computing came to picture, years back as a concept by Matei Zaharia, a Romanian-Canadian computer scientist and creator of Apache spark, a service then designed to process voluminous data.

However, the modern needs of data go beyond typical relational databases to analyzing data for streaming systems, data lakes and data warehouses. Realizing this requirement Databricks was built over the existing, open-source Apache spark core. This was later integrated with Azure and resulted in many additional improvements like auto-scaling,support for multi-language notebooks (including Python, Scala, R and Java) with built in machine learning frameworks and integrated security. The birth of Azure databricks in the market has now made it one of the most in demand and sought-after technology in terms of business analytics that promises serious results for enterprise decisions.

To the core Azure Databrick offers databricks runtime and integrated environment with additional advanced features like the role based access control (or RBAC) which includes access allocation of jobs with the collaborating team, JDBC and ODBS endpoint authentication, RStudio integration and audit logs.

So finally with Azure Databricks you get to enjoy

- Fully managed spark clusters hosted on azure
- Databrick workspace for collaboration with notebooks with multi language support (R, Scala, Java and Python)

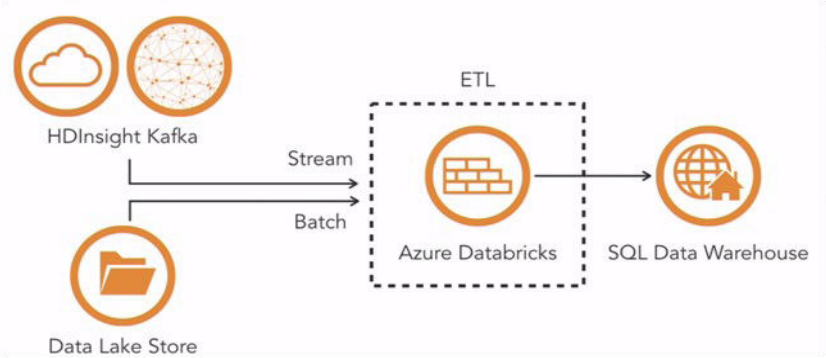
"Data is the new science; Big Data holds your answers"

- Enterprise security including Azure directory integration (authentication and authorization service)
- Integration with Azure services like SQL data warehouse, CosmosDb or power BI that keep your application in sync with other services.

Now that we have a crisp idea of what workloads Azure Databricks can handle let's understand with a practical example the hidden world of cloud technology that most enterprises use as a framework to handle their application's scalability

Architecture study of Application design

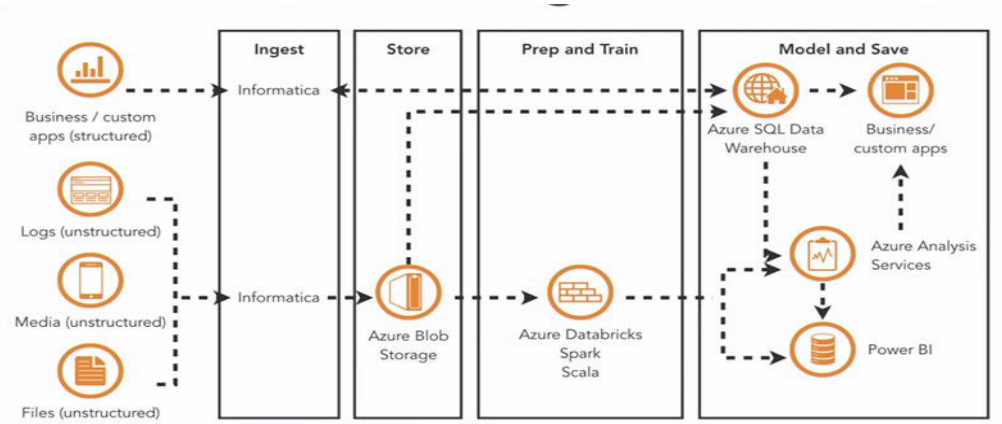
The entire process of scaling any ML application begins with loading raw data as a data stream from Kafka or as a data batch from the Data Lake store to Azure databricks that will perform the major ETL function (extract, transform and load). As databricks works over spark this further job can easily be run parallelly for any sort of data preprocessing exactly the same way mentioned in the characteristics of pipeline section. Using azure gets us the edge of processing real time voluminous and massive data. This is the most fundamental architecture and provides a clear knowledge of how a data engineer's background tasks operate.



However, if we have data coming from multiple sources, either structured or unstructured, the architecture may be adjusted so that the data is distributed over a spark cluster, enabling us to perform the exact ETL process at scale with regard to effective use of time, cost, and compute engine.

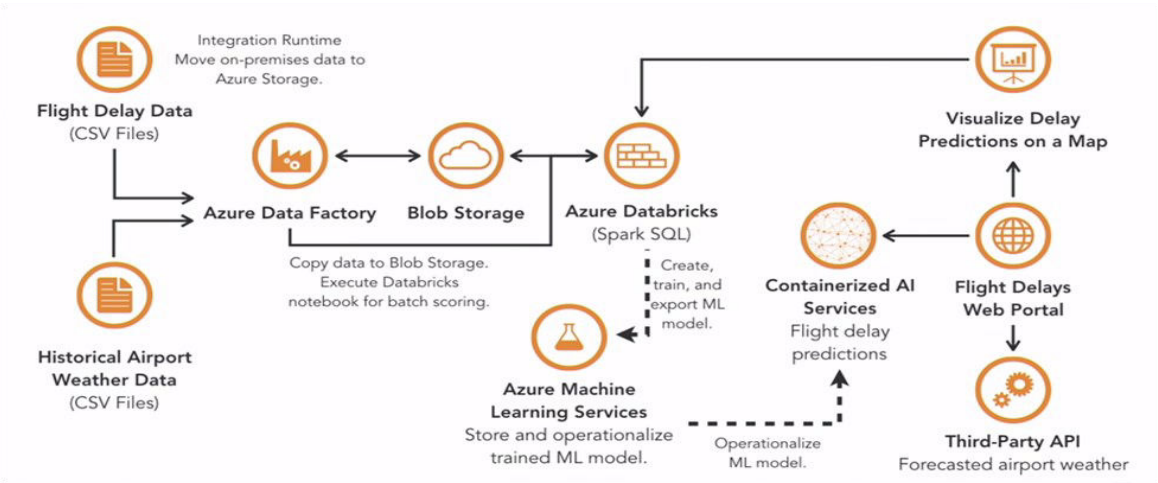
Now let's think about creating a complete application that ingests data from the real world, transforms it into a csv format, and transmits it to the Azure backend as stream or batch size data. With Azure DataFactory, an ETL service that extracts, transforms, and loads the data in a more understandable way, the path of analytical business now begins.

"Data is the new science; Big Data holds your answers"



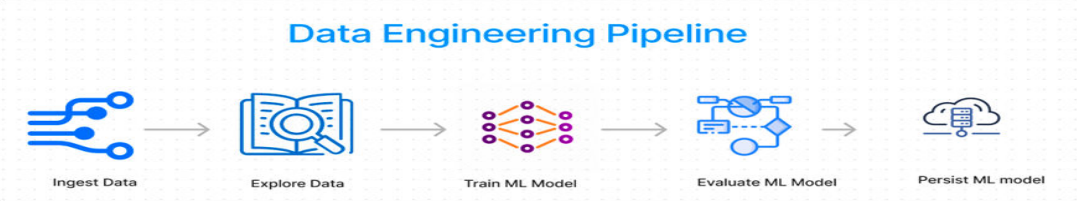
Data that moves forward, is copied to the blob storage, and is subsequently incorporated as a table in Azure Databricks. Once this data is accessible via databricks studio, we have the freedom to carry out any kind of operations using Spark SQL.

Now that the model has been operationalized and containerized as AI services, data may be expanded to Azure machine learning services. It is at this stage of development that effective recommendation and prediction systems are put into use, providing applications with strong performance. Containerized services can be made available as API services, which most businesses do as part of their business strategy for sale as third-party commercial API's.



Summarizing it down, a Data engineering pipeline for Machine Learning in Azure condenses down to the following steps

"Data is the new science; Big Data holds your answers"



Some Concluding points

As of 2022, Azure-Databricks workspace provides flexibility not only with Machine learning but also with Data Science Engineering, and SQL workloads. To the end now let's summarize a general understanding of its end-to-end architecture design, right from preprocessing steps to analyzing and visualizing the data, by skimming through the characteristics of the entire pipeline.

Pipelines	Cleansing Processing
Enterprise features	Security Monitoring
Azure integration	Other storage services Visualizations
Cost control	Azure integration

1. Involves preprocessing of data - (cleaning nulls, compressing and partitioning data)
2. Evaluating the data for its quality - by visuals, outlier inspection
3. Standard compute - SQL queries for business analytics (statistical operations on Data)
4. Complex compute - Performing ML jobs (inferencing, training models, tuning hyperparameters and accuracy of models). When the data is massive the compute time may grow exponentially, however this can be adjusted with a robust compute flavor in the workspace.
5. Serve result- store it in database and visually present it

So, to conclude

I have covered a lot of in depth and complex topics that might not get right across at first however I encourage you to keep diving into the intensity of cloud. Here are some of the best resources that are worth going through if you consider exploring ahead

Learning Resources

1. https://learn.microsoft.com/en-us/azure/databricks/
2. https://github.com/MicrosoftDocs
3. https://learn.microsoft.com/en-us/azure/databricks/notebooks/?WT.mc_id=Portal-Microsoft_Azure_Databricks