ITT/GD: Hybrid DataLake on Containers

Posted by Himansu Panigrahy Oct 25, 2019

Data Lake:

With advancement of data Analytics, Deep Learning, AI & IoT, every industry is extensively focused on extracting insights out of data for business optimization. Considering the fact that, we are in the brink of a paradigm change fueled by the explosion of data (as statistically presumed that by 2025, global data sphere would grow more than 5 times of its current size), Data Lake plays a extremely crucial role in facilitating further journey of data. Data Lake can be considered as an unlimited, unconditional storage of raw data in its native format irrespective of its size & structure. The best comparison is with the natural lake, which can be fed by multiple streams and at any time a new stream can be added without disturbing the previous setup.

Features of Data Lake:

- The main objective of Data Lake is to welcome any source system to accumulate its data at single place to perform analytics at a later stage.
- It gives unrefined insight of metadata information to data scientists.
- Brings Business Agility, as it offers flexibility & scalability at an effective cost, also accommodates any change without significant alteration to infrastructure.

- Data stored at its low label of details that brings "Data Granularity"
- Opens a wider range of operational opportunities Metadata Extraction, Data Exploration/Discovery, Data Governance & Management, Auditing & also permits Data Lineage. Complete Data Life Cycle management can be planned

Hybrid Data Lake:

Gradually, with due course of time, despite of unrevealed risks of Cloud, Industry embraced the Hybrid architecture i.e a blend of On-premise & Cloud platform, which evidently materialized the idea of Hybrid Data Lake. Primarily cloud enables better flexibility to expand indefinitely without any scalability constraints. Once the concerns around control over system, costs, availability, overall security has been shorted out, Hybrid architecture adoption went global.

Advantages of Hybrid Data Lake:

- **Business Continuity:** It enables easy configuration for High Availability of service, Data consistency, Data Backup, Disaster Management, avoid data loss
- **Innovation:** Opens up door for multiple tools that enable smooth
- **Improves Flexibility & Scalability:** availability of unlimited resources on-demand
- **Simplifies Data Ingestion:** Invites multiple sources with varieties of data
- **Cost effective:** It offers economic cloud storage & any resource can be availed conveniently
- UI based **One-click** Management of Resources/Jobs/Applications
- Cloud offers numerous **APIs/Inbuilt tools** to perform Data Analytics/Machine Learning on top of data

Next-Gen Data Lake: (Hybrid Data Lake on Containers)

How is the thought of taking whole Data Lake, or better to say Hybrid Data Lake itself to Containers !!

The idea might sound fancy & intangible. I would not deny that the feeling was mutual few months back, when I worked on a proposal of building a complete platform over containers rather than running a micro-service, which was beyond the reason of containers origin. However now the containerization is no more a concept, it has advanced enough to implement complex architectures way-ahead than just micro-service.

Lets take a step back to understand concept of containers !!

Current era of technology has already surpassed the phase of disintegration of Monolithic system into micro-services, which led container predominantly adopted. Even Gmail, Youtube & Google Search applications are now running on containers, because of the astounding features it provide.

- **Containers** allow application sand-boxing, which are entirely decoupled from the underlying infrastructure & OS, efficiently light weight, with its own file-system, resources (cpu/vcores, memory, network), process space, capable to running without effecting other containers, easily portable across anywhere regardless of the target hosting system either in local laptop/bare-metal server/cloud without any extra effort to make it compatible to the underlying system.
- **Docker & Docker-Swarm** made the containerization, clustering, scheduling & management of containers effortless.
- Due to its light weight, it can be deployed at high density with abundance of workload. However it was not yet ready for productionization due complexity of coordination, networking & lack of proper governance, causing the system disruptive.
- Then **Kubernetes** emerged as an unified solution to all these complexities, which itself is a comprehensive system made Governance & orchestration of containers easy as pie.
- It extends features to achieve HA(High Availability), Scalability & Data Persistance, Inbuilt Security features (sa, rbac, ssl, secrets, token etc.), Networking, Ingress/Egress policies, CRDs (Custom Resource Defination) to design our own custom application deployment, Custom Operators etc. Briefly speaking Kubernetes alleviates Private Cloud concept & PaaS (Platform As A Service) offering.

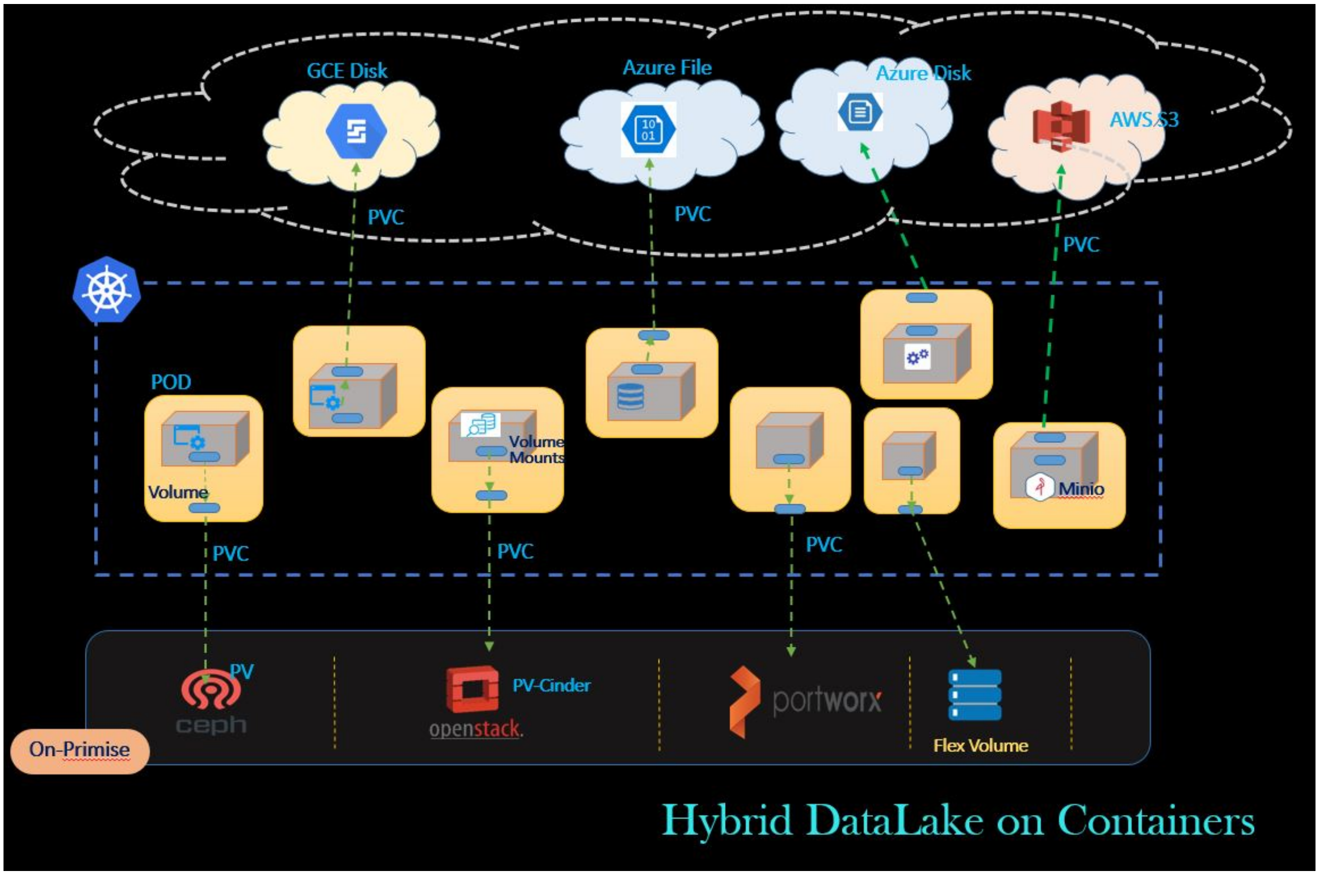
Hybrid Data Lake on Containers !!

Data Lake is all about hassle-free welcoming of diverse Data Sources. When we take the Data Lake to containers, it acts as combo of Public-Private cloud.

- A layer of containers in between Bare-metal/Cloud could storage, could take the Hybrid Data Lake to a next label. Especially the Persistent storage on cloud can be accessed through APIs/third party tools deployed in containerized applications/services. eg: Minio is an open source tool can be provisioned to access S3 storage on AWS
- This could elevate the complexity to some extend, however with Kubernetes, it is feasible to attain.
- Even various opensource tools available in market, are equipped enough to achieve it.
- Kubernetes supports various types of persistent storage solutions either in Cloud/On-prim bare-metal such as (GCE-Persistent Disk, AWS S3, Azure Disk/File, Cinder, glusterFS, rbd, portworx, Flex-Volume).

Instead of elaborating much, Let's look at the architecture which would clarify further.

# The concept is open for discussion. Please provide your feedback & connect with me to take this forward !!



Special Thanks to Hemant Dindi Rachit Mathur for helping me out !!
76 Views Tags: cloud, container, data, kubernates, datalake, itt-gd, #gdtechblogathon

Vinod Sundaram
Oct 25, 2019 12:30 PM
Hybrid DataLake on containers.... Interesting concept!!!! Thanks for giving this insight

Faisal Sait
Oct 25, 2019 12:01 PM
That is great insight, well written.

Pradeep Hosakoti

Oct 25, 2019 11:49 AM

Interesting topic, all the very best to the team..!