

**QUESTION:**

How did you determine that a Data Lake is the right technology for the enterprise?

ANSWER:

Data Lakes are ubiquitous with storing data for the "3 Vs" which are Volume, Velocity and Variety. At Microsoft, we have huge data volumes, data coming from traditional operational systems and streaming sources, in many different formats. It is very helpful to be able to ingest this data into a foundation like a data lake without worrying if we have enough space, or without the need to first model the data prior to storage in a traditional database. In some cases, the data is unstructured and only fits well in a data lake. Azure Data Lake generation 2 is a great technology to store any data for analytical purposes regardless of which of the "Vs" you are focused on in your scenario.

QUESTION:

What do you mean when you say "data products"?

ANSWER:

A data product is a data asset that enables the generation of insights and intelligence from data. Examples of data products could be; Machine Learning Models, Connected/Aggregated foundation datasets, analytics models, reports and dashboards. Data Products also include the Data APIs that enable generating the insights and intelligence to user facing applications.

QUESTION:

What are some of the data services you have set up in your Enterprise Data Lake?

ANSWER:

Our Enterprise Data Lake is an Enterprise Data Management Platform for our Analytics, Data Science (ML/AI), and big data processing applications. Services built around and into our enterprise data lake include data ingestion services, data standardization services (common analytical models, data unification, PII (Personal Identifiable Information) de-identification, Merging etc), data quality management, data access management, data compliance (GDPR, SOX etc), data governance services, and data catalog publishing.

QUESTION:

What are some of the ways you have accelerated your delivery time of data products??

ANSWER:

Establishing a shared foundation for trusted and connected data as a single destination to discover and gain access to enterprise wide data, and democratizing secure and compliant access to the data via modalities that widely used by data practitioners (data analysts, data scientists, data engineers, and data savvy business users), and automating the DevOps capabilities to operationalize data products as high scale services, are the foundations for accelerating our data products delivery.

QUESTION:

Do data consumers need to learn new skills to use data stored in the data lake?

ANSWER:

Maybe? If your background is ONLY in SQL, then you have several options in Azure Data Lake.

First, you can use Azure Synapse Analytics to directly query data in the lake exposed as a view in your Synapse workspace. This is very helpful for some scenarios where sub-second response is not required.

Second, it is still extremely common for a data warehouse or other reporting solution to be fed from the data lake, so in this scenario, there is no impact to consumers, only some rewiring by engineers building and maintaining the reporting solution.

Finally, if you have a need to directly interact with data stored in the lake, it is useful to learn some big data skills like Python/Scala in Notebooks to self-serve against the data stored in the lake. This approach also honors a critical modern principal of bringing compute to the lake to avoid unnecessary proliferation and copying of data outside of the data lake.

QUESTION:

How do you know that you've improved the quality of your data?

ANSWER:

We have integrated Data Quality (DQ) Management capabilities in our Enterprise Data Lake with descriptive DQ rules and probabilistic ML models that assess data quality against a broad set of trusted data criteria and general data quality dimensions (accuracy, completeness, latency etc). These integrated services enable us to gain "always on" insights on the quality of our data and take fast turn corrective actions both proactively and reactively to address.

QUESTION:

What does it mean to democratize data?

ANSWER:

Democratizing data means making data accessible to every function, team, and person in an organization who has an opportunity to use and apply data, to enable/deliver a differentiated value outcome. Our motto is "Responsible Data Democratization", which is striking a balance between enabling access and having the required controls in place for secure, compliant, and governed applications of data.

QUESTION:

Are self-service reporting capabilities a good idea for enterprises?

ANSWER:

Absolutely! It just needs to be done responsibly. There should be a reasonable business justification for access to data approved by the owner or steward of the data. Ideally, there should be no standing permissions for data access. Finally, compute should be brought to the data whenever possible to minimize data copying outside of the lake where governance controls may or may not be

present. These are all critical parts of our data strategy at Microsoft. If done responsibly in a modern fashion as described, self-service helps to maintain needed business agility.

For more information

Microsoft IT Showcase

microsoft.com/itshowcase

© 2020 Microsoft Corporation. This document is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY. The names of actual companies and products mentioned herein may be the trademarks of their respective owners.