10 BEST PRACTICES FOR ARCHITECTING DATA LAKE SOLUTION

Adam Wiszniewski, 16 May 2019

softserve

A few words about me

- Senior Big Data Engineer at SoftServe
- I design data pipelines and data lakes
- Cloud enthusiast: AWS and GCP stacks
- Experienced in banking, airline, FMCG, marketing data domains
- Analytical mindset

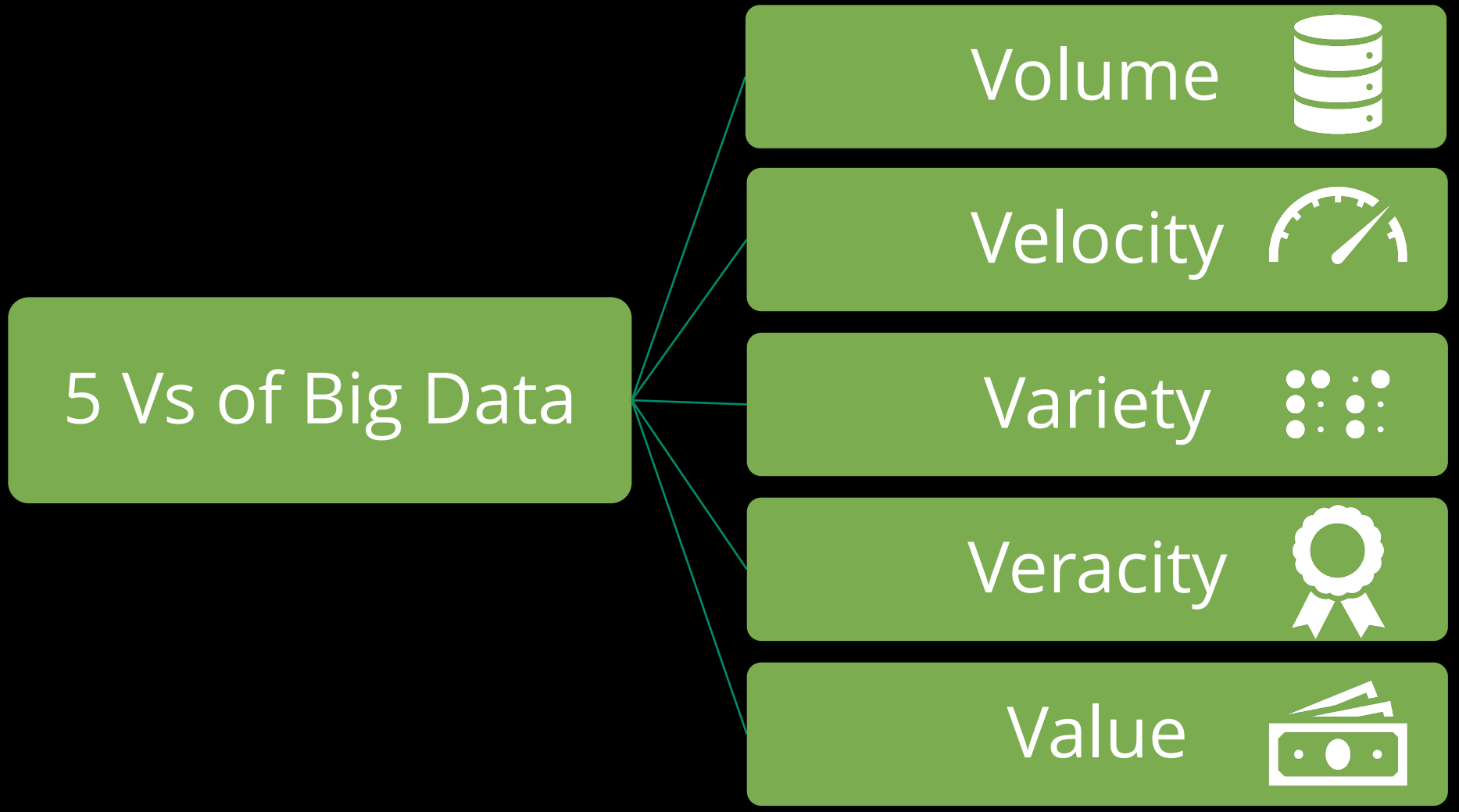
softserve

Agenda

- How to define Big Data
- Data Lake intro
- Data lake best practices
- Conclusions
- Q&A and place for sharing yours best practices!

softserve

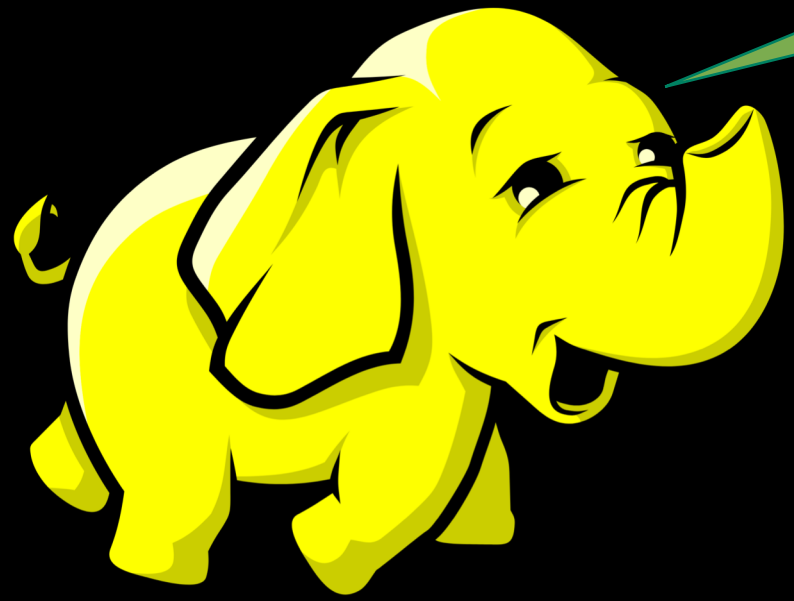
How to define Big Data?



softserve

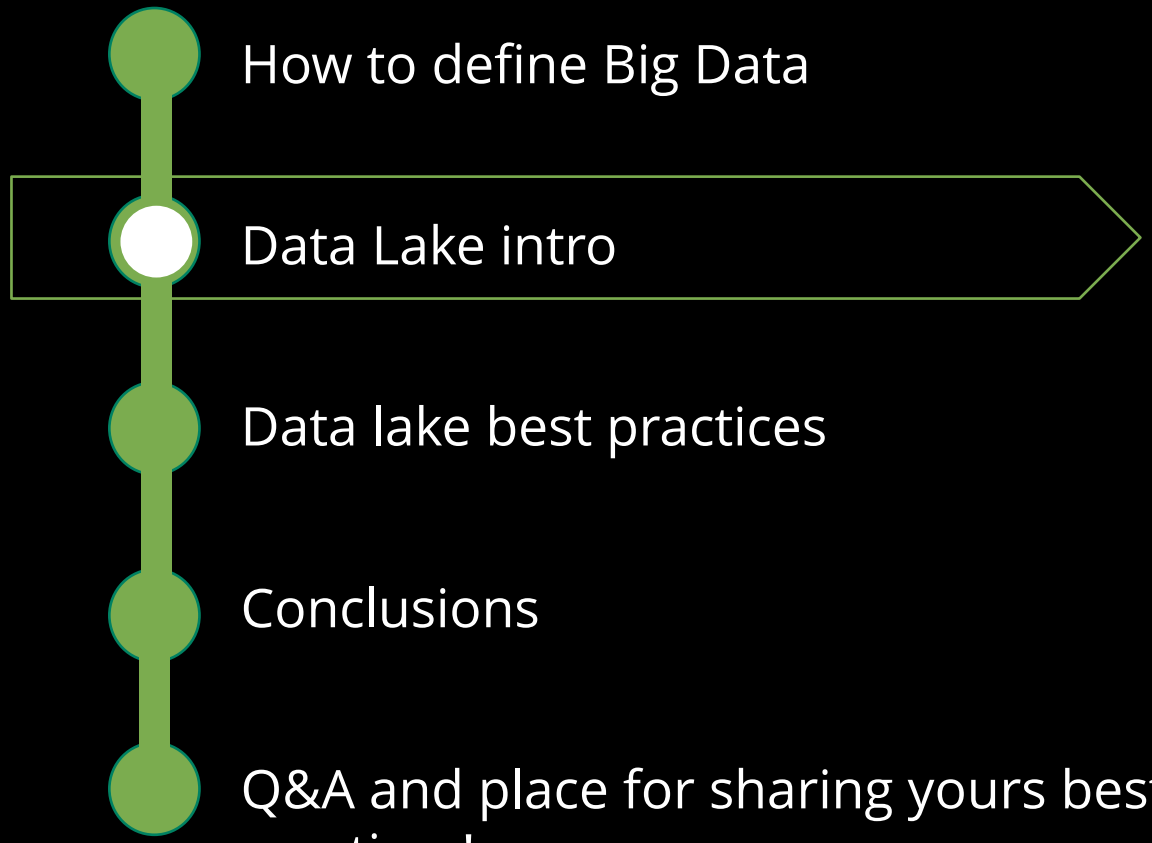
Big Data maturity

I'm thirteen!
Don’t treat me as promising technology!



I’m almost 5!
Stop calling me new technology.

softserve



How to define Big Data

Data Lake intro

Data lake best practices

Conclusions

Q&A and place for sharing yours best practices!

softserve

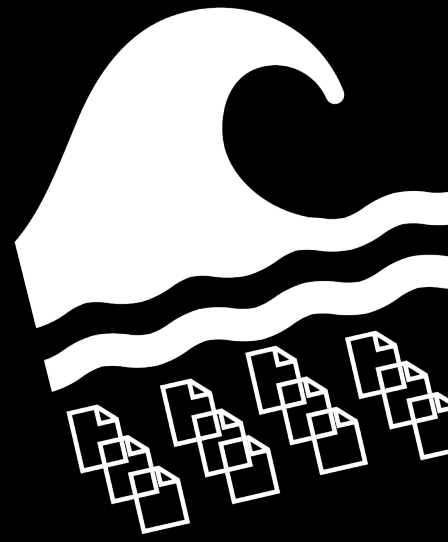
What is data lake?

- Centralized data repository
- Democratized data among company
- Usually golden source of data for data scientists
- In most cases schema on read
- Stored on HDFS or cloud systems (S3, Azure Data Lake, GCP CloudStorage)

softserve

Data Lake usage

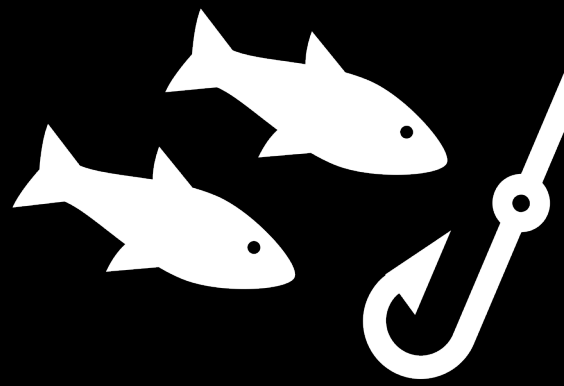
- Central data repository
- Candidate for central data platform
- Source for many data marts
- Place to store all object types
- Playground for data science and R&D teams
- Generate insights from different sources



softserve

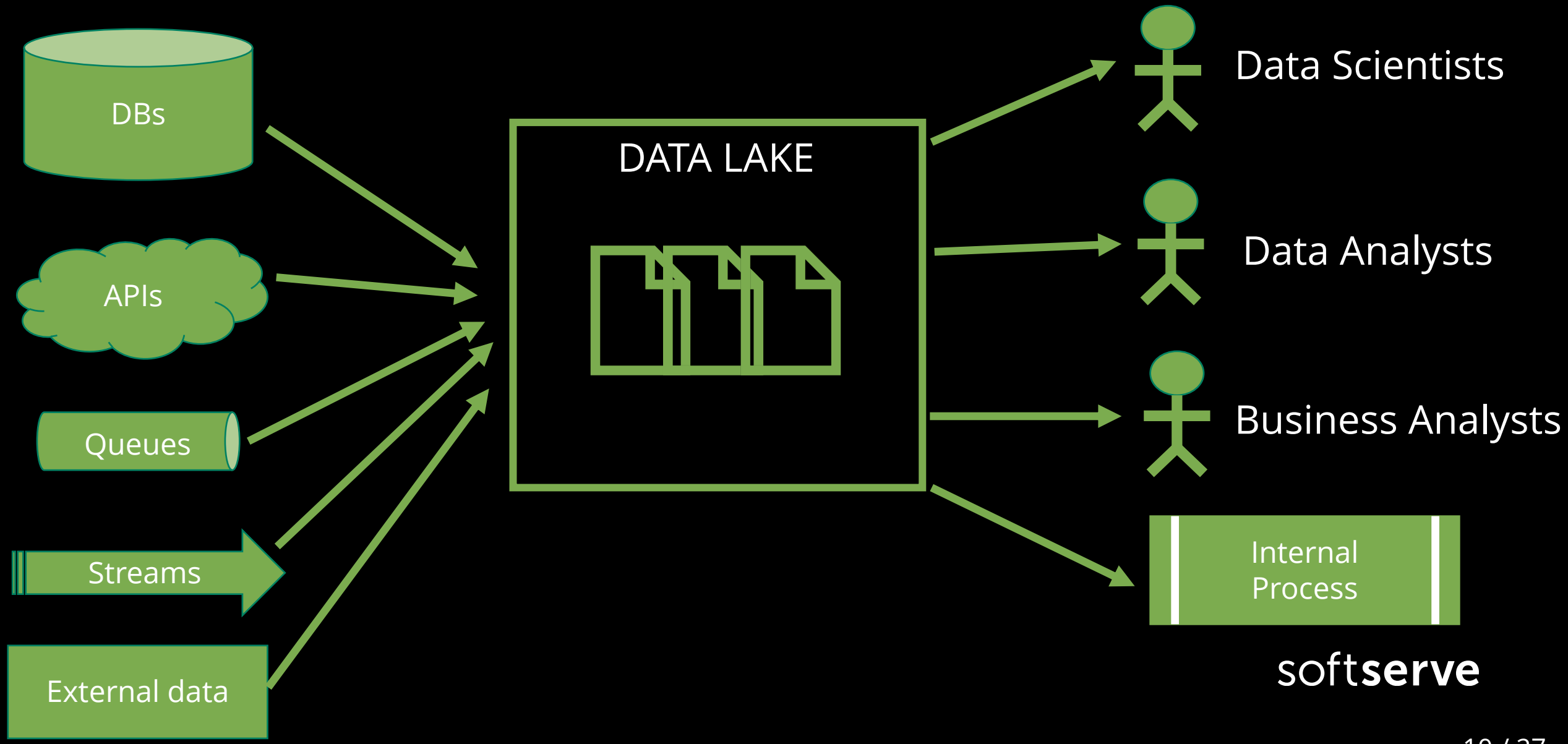
Data Lake concerns

- Skills
- Costs
- Platform
- Data security
- Data Lifecycle
- Data Governance

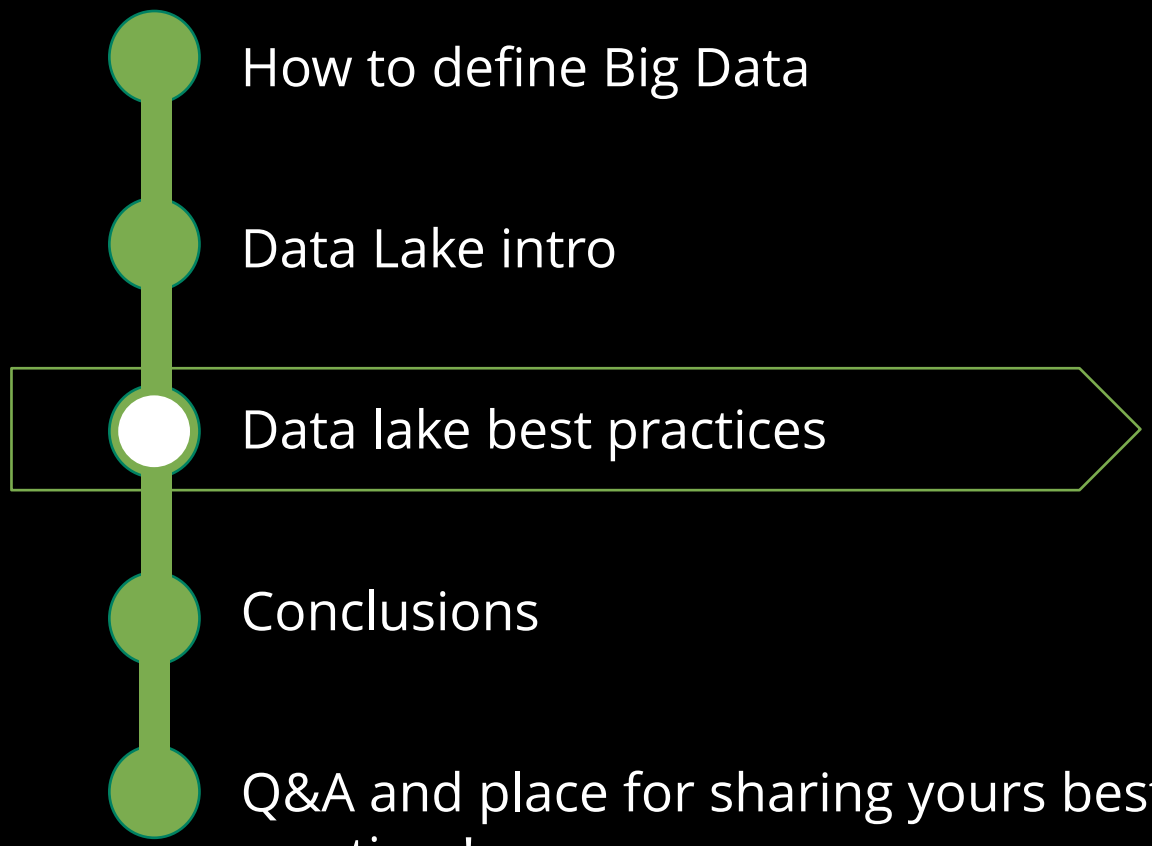


softserve

Generic Data Lake Architecture



softserve


How to define Big Data
Data Lake intro
Data lake best practices
Conclusions
Q&A and place for sharing yours best practices!

softserve

BEST PRACTICES

#01 - Data on-boarding patterns
#02 - Preservation of source data
#03 - Data tiering
#04 - Adjust file sizes
#05 - Watch out for data silos

#06 - Monitor data lake pipelines
#07 – Data lifecycle management
#08 - Look after data quality
#09 - Metadata repository
#10 - Enable self-service discovery
#11 - Maintain security standards

softserve

#01 - Data on-boarding patterns

- Define generic principles for onboarding raw data
- Partition data at least by time
- Map and audit all ingestion pipelines
- DMZ bucket for external ingestion
- On raw bucket keep data from different sources in separate top keys

Raw source #1

Raw source #2

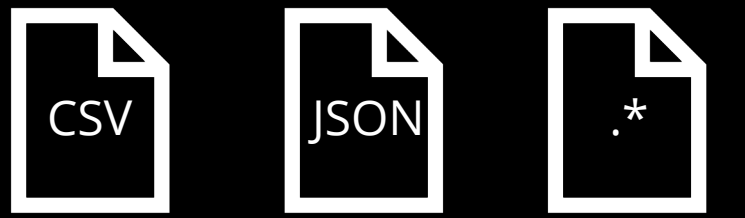
Raw source #3

softserve

#02 - Preservation of source data

- Keep untouched original raw data ideally with no transformations applied
- It will be treated as golden source in case of reprocessing needs
- Data should follow natural ingest partitioning scheme

```
0000000 0000 0001 0001 1010 0010 0001 0004 0128
0000010 0000 0016 0000 0028 0000 0010 0000 0020
0000020 0000 0001 0004 0000 0000 0000 0000 0000
0000030 0000 0000 0000 0010 0000 0000 0000 0204
0000040 0004 8384 0084 c7c8 00c8 4748 0048 e8e9
0000050 00e9 6a69 0069 a8a9 00a9 2828 0028 fdfc
0000060 00fc 1819 0019 9898 0098 d9d8 00d8 5857
0000070 0057 7b7a 007a bab9 00b9 3a3c 003c 8888
0000080 8888 8888 8888 8888 288e be88 8888 8888
0000090 3b83 5788 8888 8888 7667 778e 8828 8888
00000a0 d61f 7abd 8818 8888 467c 585f 8814 8188
00000b0 8b06 e8f7 88aa 8388 8b3b 88f3 88bd e988
00000c0 8a18 880c e841 c988 b328 6871 688e 958b
00000d0 a948 5862 5884 7e81 3788 1ab4 5a84 3eec
00000e0 3d86 dcb8 5cbb 8888 8888 8888 8888 8888
00000f0 8888 8888 8888 8888 8888 8888 8888 0000
0000100 0000 0000 0000 0000 0000 0000 0000 0000
*
0000130 0000 0000 0000 0000 0000 0000 0000
000013e
```



softserve

#03 - Data tiering

Raw bucket
original format

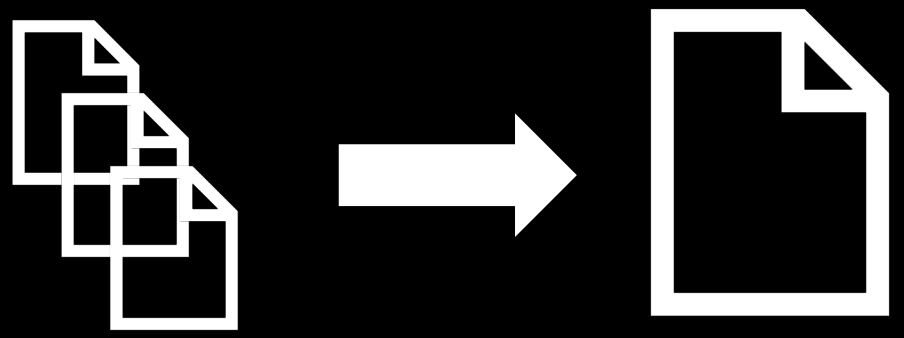
Discovery bucket
Transformed data
Sandbox

Data Mart bucket
Specialized data

softserve

#04 - Adjust file sizes

- Avoid small file sizes. Reasonable size is 256MB or greater
- Small files makes pressure on S3 API/ nameNode during data consumption
- If it is possible implement data compaction mechanism



softserve

#05 - Watch out for data silos

- Think about data strategy
- Don’t onboard data in ASAP-mode, everything in data lake should be connected with wide Data Strategy
- Pushing everything to DL causes redundancy, issues, confusion and costs.

softserve

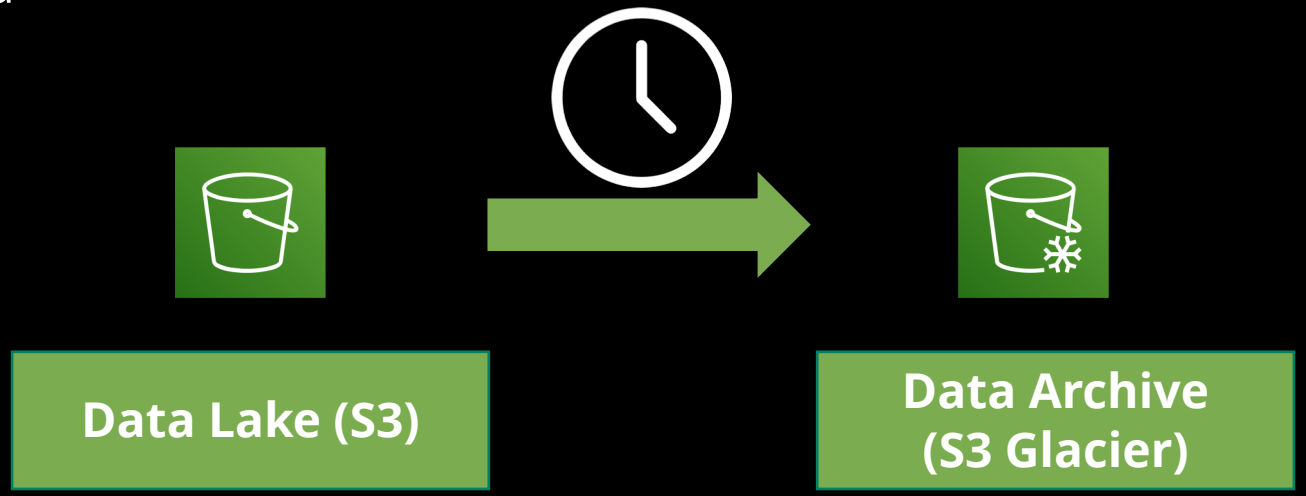
#06 - Monitor data lake pipelines

- Introduce yours domain-specific metrics and save them in AWS CloudWatch
- Build dashboard in Grafana/CloudWatch
- Setup some alarms for unnormal behavior i.e. “no data ingested in last 6h”
- Keep eye on it!

softserve

#07 – Data lifecycle management

- Declutter data lake and retire old partitions to archival store
- Data lifecycle policies should be defined within organization
- It reduces storage costs
- Retention policies are under GDPR



softserve

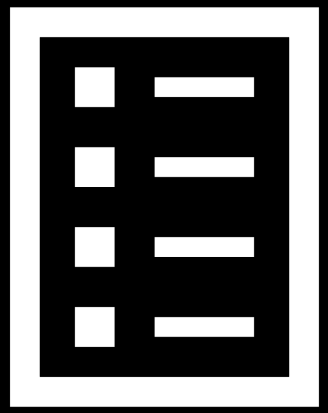
#08 - Look after data quality

- Don’t consider your job in terms succeed / failed
- Introduce post-processing validation checks
- Setup automatic job which will perform data quality checks like:
  - Size of partitions
  - Count of records
  - Business quality checks on data
- Notify about data quality status

softserve

#09 - Metadata repository

- Map your DL assets to catalog for better transparency
- For tabular data consider usage of AWS Glue Data Catalog as serverless hive-compatible metastore
- For raw data, like schema-less and blobs could be Elasticsearch or RDS for metadata storage
- Good practice is having business data catalog, with description of datasets for BAs/PMs/DS



softserve

#10 - Enable self-service discovery

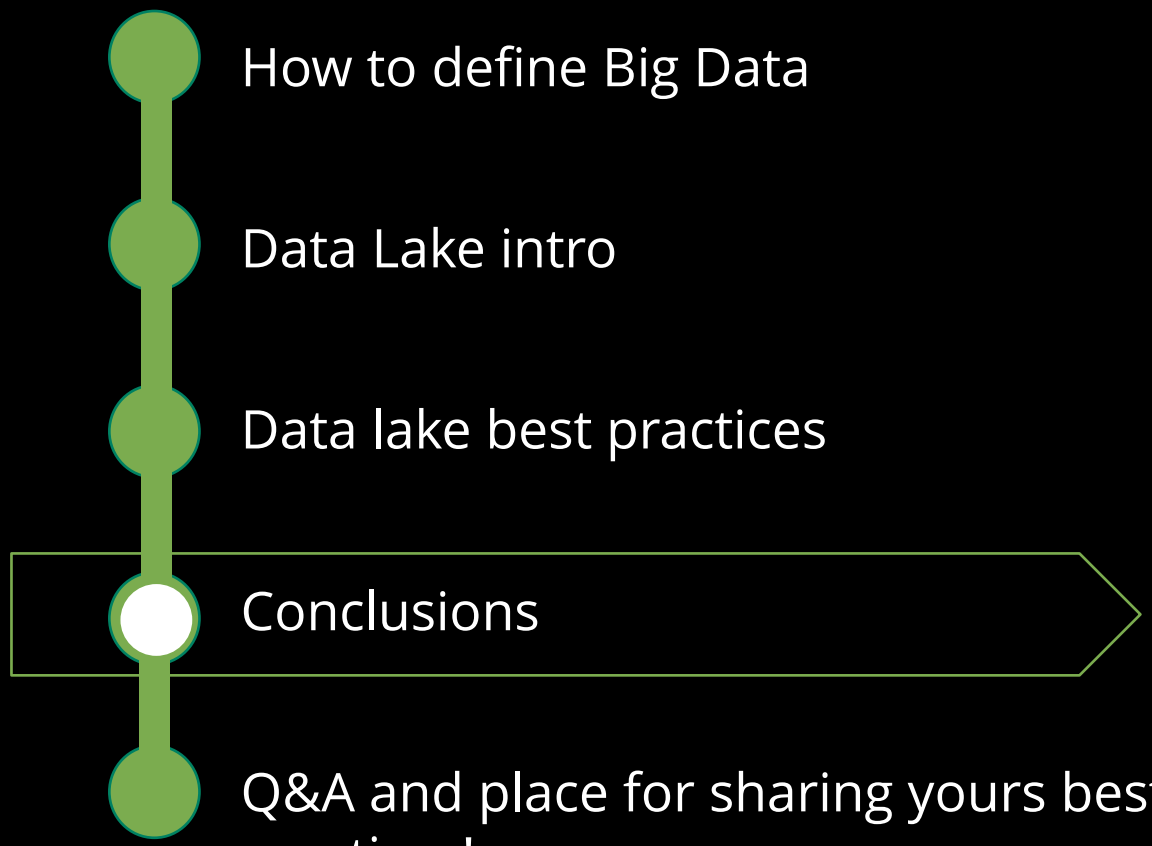
- Give users access to explore DL - ideally using SQL-like language
- For Parquet data stored in higher tiers Athena or EMR/Presto engine works nice
- Educate your users about usage scenarios and integration options
- Provide manuals to integrate with Machine Learning supporting tools like SageMaker or Zeppelin
- Having good metadata repository and self-service platform gives a real boost to Data Scientists and Data Analysts

softserve

#11 - Maintain security standards

- Create IAM roles for different user groups or datasets with S3 access restrictions
- If possible integrate it with Active Directory
- You could also restrict access on metastore level
- For highly regulated domains - S3 objects might have tagging for determining IAM access rights i.e. “SecurityLevel=Restricted” for certain data

softserve



- How to define Big Data
- Data Lake intro
- Data lake best practices
- Conclusions
- Q&A and place for sharing yours best practices!

softserve

Conclusions

- Don’t advertise or treat Data Lake as enterprise silver bullet
- Think about data strategy
- Keep eye on data pipelines
- Democratize data assets across interested parties
- Have a clear understanding what data is inside DL
- I did my best to extract general patterns, but take in mind that best practices may differ from scenario to scenario

softserve

THANK YOU!

softserve

Q&A and yours best practices

Q&A

Yours Best Practices

Next
Exit