Serverless Architecture for Data Science Projects

Roman Golovnya
12 December 2020

About me

- Data Engineer @ ResMed Group
- Roman has a decade of experience gained in Data, IT and Financial Services
- Postgraduate qualifications in Cloud Computing, Data Analytics and Finance
- Founder and organiser of Data Science Engineering Club
- roman.golovnya@gmail.com

Agenda

- Introduction to serverless
- Challenges in Data Science projects architecture angle
- AWS serverless services
- Suggested serverless services Data Science projects
- Serverless for Machine Learning
- Hands-on resources
- Summary
- Q&A

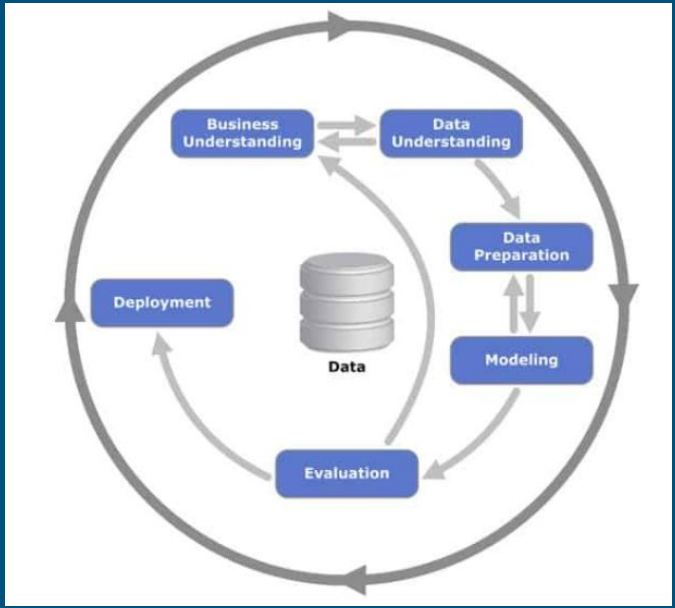
1 - Introduction to serverless

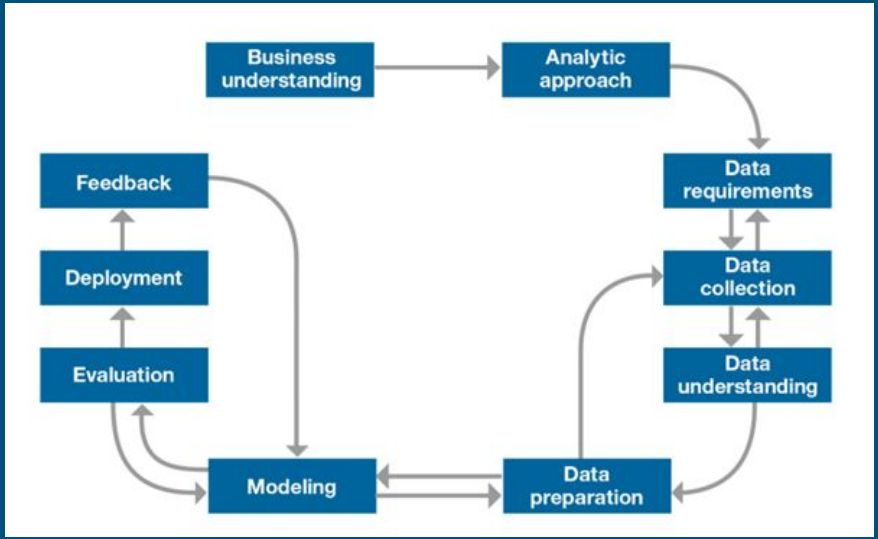
- **A serverless architecture is a way to build and run applications and services without having to manage infrastructure.**
- **Serverless architectures are application designs that incorporate third-party “Backend as a Service” (BaaS) services, and/or that include custom code run in managed, ephemeral containers on a “Functions as a Service” (FaaS) platform.**
- **https://martinfowler.com/articles/serverless.html**

- **You only pay for the time the function run**
- **You don’t manage or own the infrastructure**
- **You can expect the service to be reliable and highly available**

Data Science workflow

- **CRISP DM model updated**
- https://www.datascience-pm.com/crisp-dm-2/





2 - Challenges in Data Science projects

- **Over 80% work spend on data integration, data wrangling**
- **Heterogeneous data**
- **Vast amount of data**
- **Real-time data**
- **Security and Governance**
- **Deployment to production**
- **Cost optimization**

3 - AWS serverless services

Compute

AWS Lambda, Amazon Fargate, Amazon ECS, EKS

Application Integration

Amazon API Gateway, Amazon SQS, Amazon SNS, AWS Step Functions, Amazon EventBridge, AWS AppSync, Kinesis Firehose

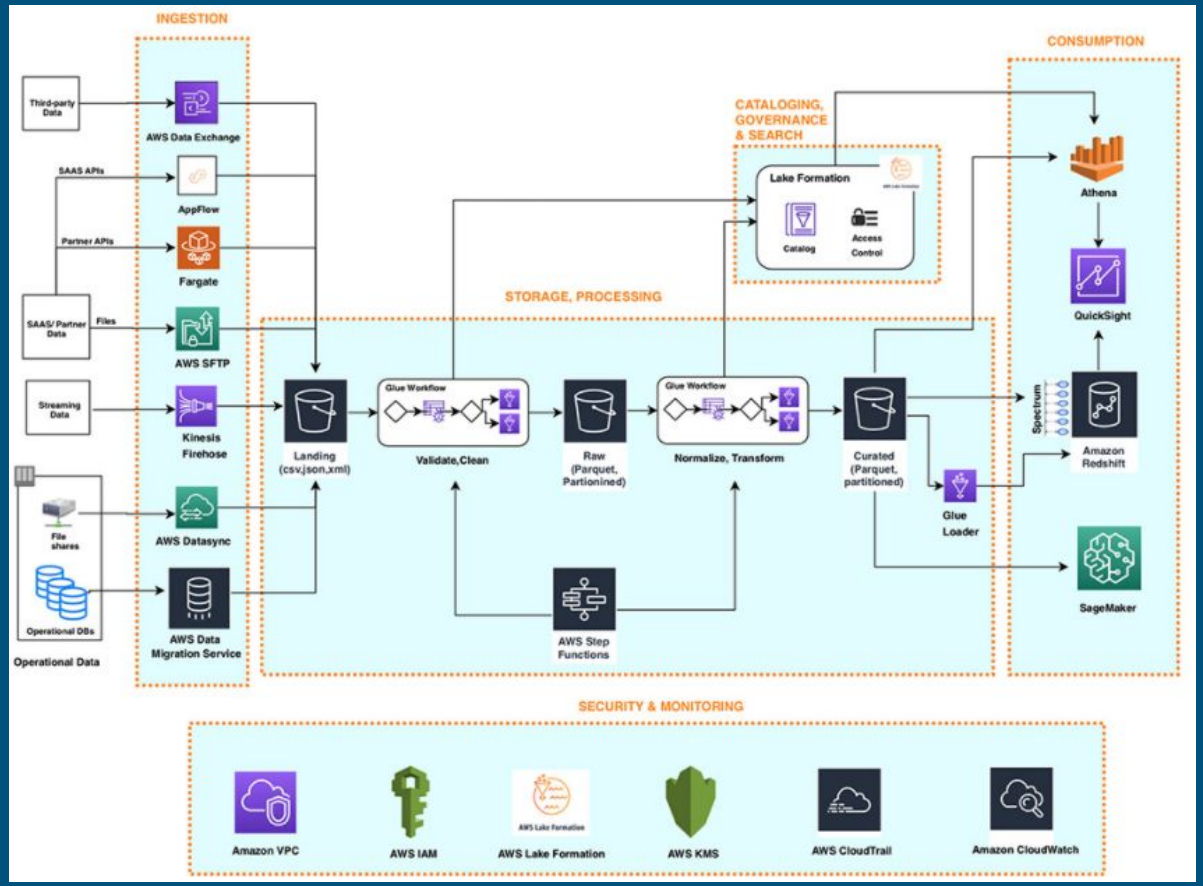
Data Store

Amazon S3, Amazon DynamoDB, Amazon RDS Proxy, Amazon Aurora Serverless

Frameworks

AWS Serverless Application Model (SAM), serverless

4 Serverless services for Data Science Project



- https://aws.amazon.com/blogs/big-data/aws-serverless-data-analytics-pipeline-reference-architecture/

5 - Serverless for Machine Learning

- **Machine Learning inference with Lambda**
- **https://aws.amazon.com/blogs/compute/pay-as-you-go-machine-learning-inference-with-aws-lambda/**

- **Extend Lambda with EFS for hosting dependencies**

- **Use provisional concurrency to avoid cold starts**

- **Serverless continuous integration https://github.com/lambci**

6 - Hands-on resources

- https://github.com/aws-samples/serverless-data-analytics
- https://github.com/aws-samples/amazon-serverless-datalake-workshop
- https://github.com/aws-samples/aws-lambda-serverless-inference

4 - Conclusion

- **Leverage serverless architecture**
- **Automated high availability**
- **Flexible, automated scaling**
- **Build and deploy faster**
- **Pay when it run**
- **It will focus your efforts on what provides more value to users.**

Conclusion