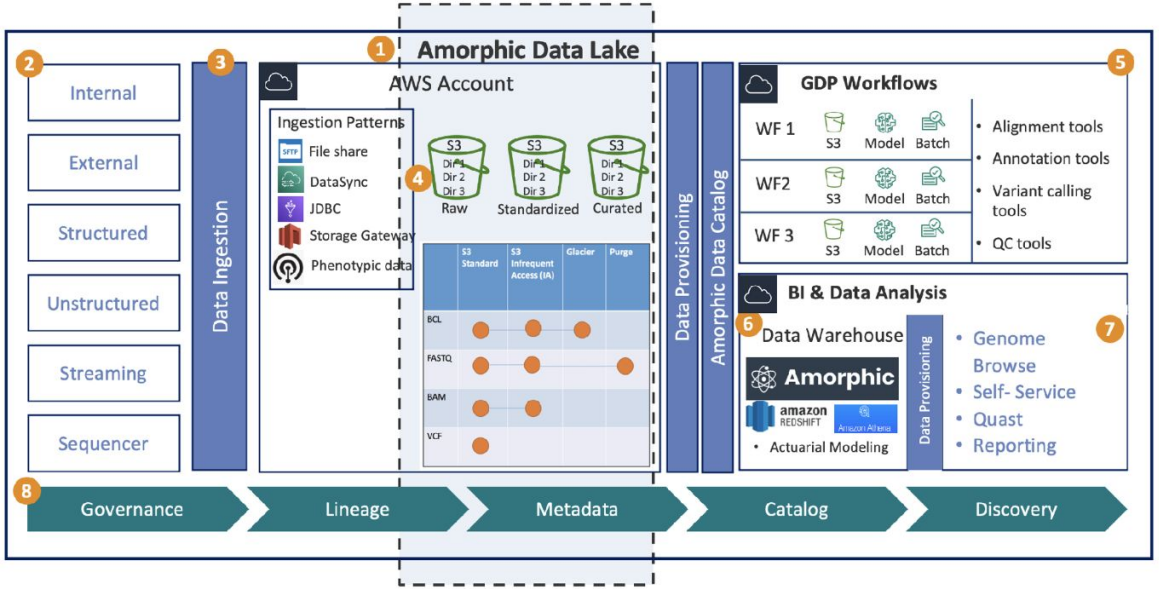
Amorphic - Genomics Data Cloud

Cloudwick

Agenda

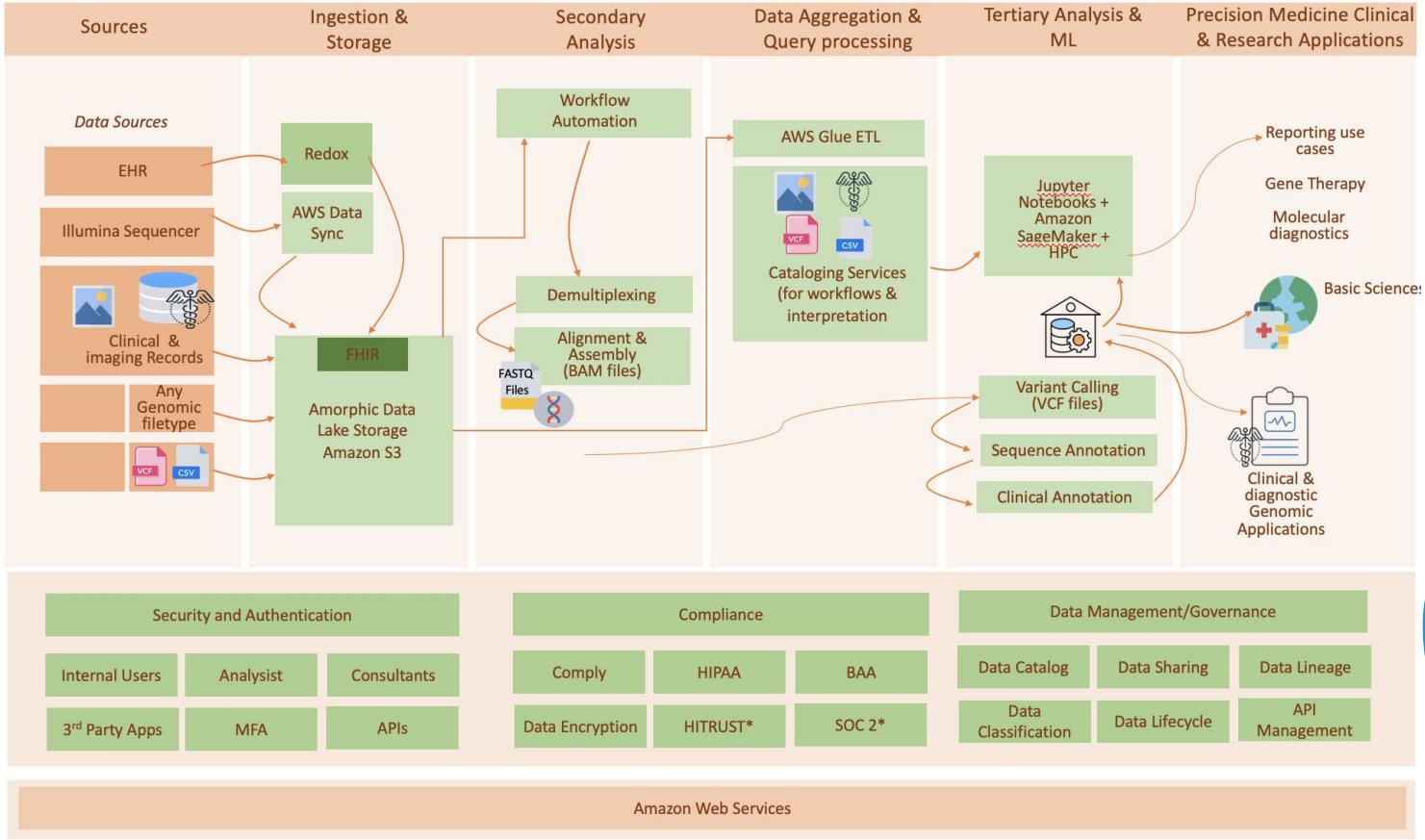
- Blueprint of Genomics Data C
- Data Workflow and Management
- Amorphic Conceptual Architecture
- Amorphic Technical Architecture
- Use Cases for Genomic Data Platform
- Genomic Reporting Use Case
- Q&A

Blueprint for Genomics Data Platform

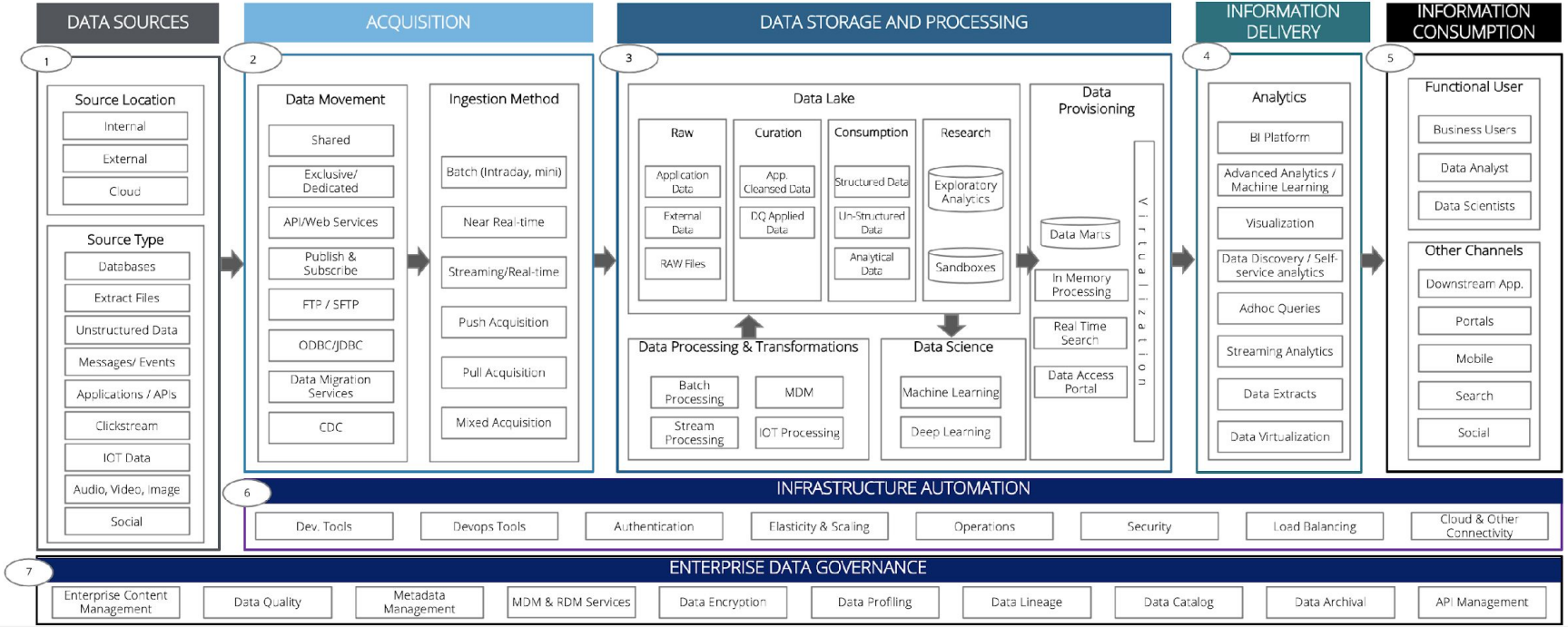


1. GDP Data Lake provides for the inclusion of all buckets and directories from all accounts including catalogs.
2. Inclusive of all potential internal and external data sources and types
3. Standardized and reusable ELT / ETL processes to ingest raw data from sources including CDC and transform per genomics rules as needed
4. The raw BCL files are read once and are rarely touched after that. The FASTQ may be accessed for some time but they can then be purged because you can regenerate them if you need to. BAM files are heavily accessed but the access patters are known to decrease once after the variant calling operation. Lastly, VCF files are small in size but highest in value as it can be used as input for many analysis processes downstream.
5. Workflows like secondary and tertiary analysis
6. Amorphic data warehouse options to leverage for analytics and BI projects
7. Data Consumption can include any BI, dashboarding, Genome browser and reporting tool
8. Enterprise data management and infrastructure automation provide common supporting functions like Metadata, MDM, Governance and DEVOPS

Data Workflow & Management



Amorphic Conceptual Reference Architecture



1. DATA SOURCES

Authoritative source of data in existing internal and external repositories

2. ACQUISITION

ELT processes to ingest raw data from sources into the Data Lake, including Change Data Capture, and prepare it for reports and dashboards and downstream consumption

3. DATA STORAGE

Data Lake to store raw, curated and processed information from different sources. Data Warehouse/ Datamart for reports and dashboards

3. DATA PROCESSING

Refinery of batch, stream, IoT, MDM processing services to support business

4. DATA DELIVERY

Delivery of information to downstream apps and portal. APIs to enable data delivery to other applications either scheduled or on ad-hoc basis

5. DATA CONSUMPTION

Dashboards and Reporting solution for end-user consumption.
Advanced Analytics and AI/ ML Data Science Models

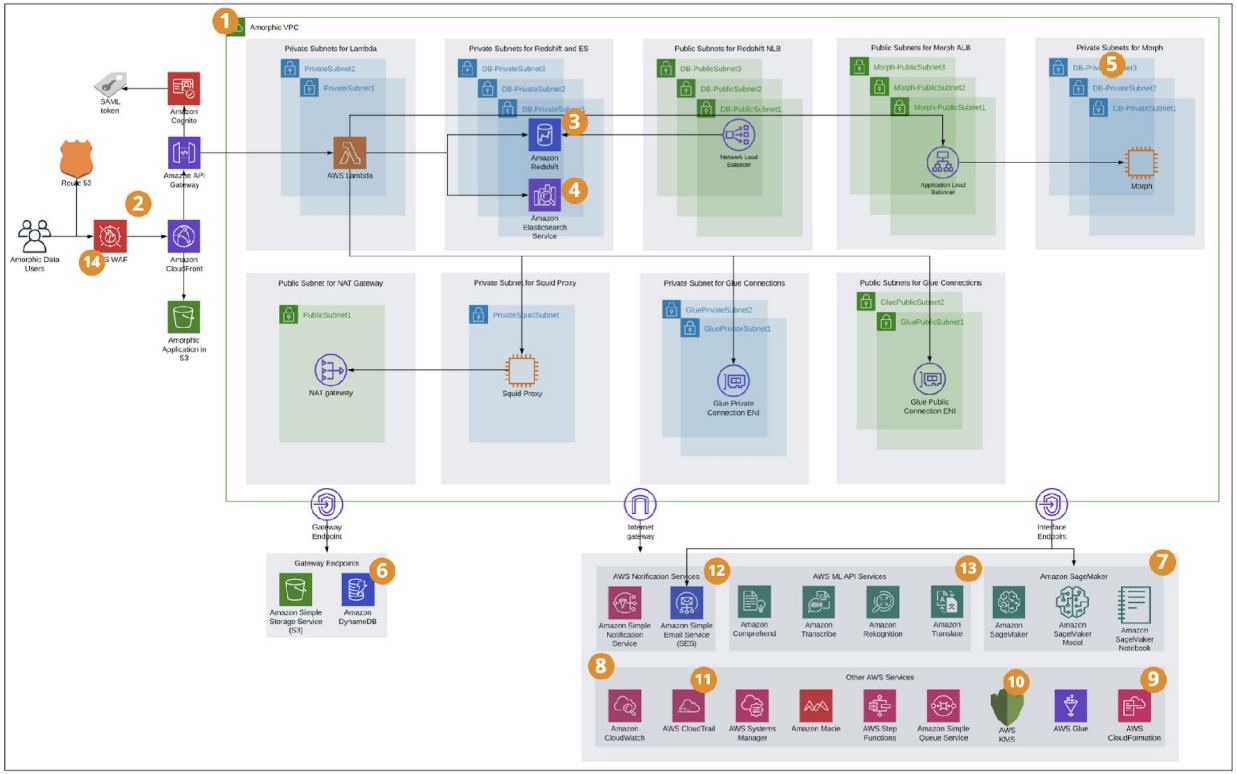
6. ENTERPRISE DATA MANAGEMENT

Common and supporting functions, like Metadata, MDM, Governance, etc.. from ingestion to delivery

7. INFRASTRUCTURE AUTOMATION

Infrastructure provisioning/ as a code and monitoring tools along with orchestration of the pipeline

Amorphic Technical Architecture



1. 100% Private deployment
2. Amorphic serverless application UI
3. Enterprise data warehouse
4. Enterprise search
5. Drag&drop ETL
6. Amorphic catalog
7. Data science and ML capability
8. Enterprise grade security
9. 90 minutes deployment using IaaC
10. Encryption at rest and in motion
11. Audit capability
12. Notifications & Alerts
13. Intelligent catalog for unstructured data
14. Amorphic application firewall

Use Cases for Genomic Data Platform

Data Transfer and Storage

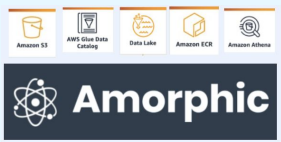
High-scale performant data ingestion from sources like DNA sequencer using Amorphic connections built on the top of AWS DataSync, Storage Gateway, DMS and AWS Glue

Workflow Automation

Amorphic's simplified orchestration for AWS Batch & ECR allows running and automating parallelizable workflows for secondary analysis

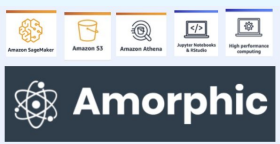
Data Governance

Amorphic catalog enables organizations to harmonize multi-omic datasets and govern robust data access controls and permissions across a global infrastructure



Deep learning for Tertiary analysis

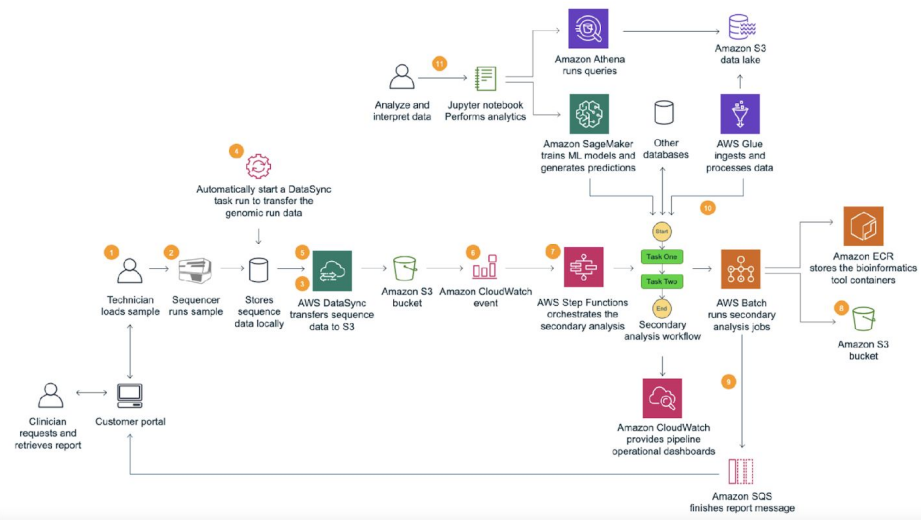
Amorphic notebook and ad-hoc query engine accelerates analysis of big genomics data by leveraging machine learning and high-performance computing.



Applications

Amorphic deep learning and ML allows organizations can establish a differentiated capability in genomics to advance their applications in precision medicine and patient practice

Use Case for Genomics Reporting

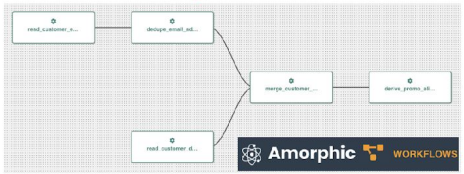


1. A technician loads a genomic sample on a sequencer.
2. The genomic sample is sequenced and written to a landing folder that is stored in a local on-premises storage system.
3. An AWS DataSync sync task is preconfigured to sync the data from the parent directory of the landing folder on on-premises storage, to a data set in Amorphic.
4. A run completion tracker script running as a cron job, starts a DataSync task run to transfer the run data to an Amazon S3 bucket. An inclusion filter can be used when running a DataSync task run, to only include a given run folder.
5. 5. DataSync transfers the data to Amazon S3.

6. An Amazon CloudWatch Events is raised that uses an Amazon CloudWatch rule to launch an AWS Step Functions state machine.
7. The state machine orchestrates secondary analysis and report generation tools which run in Docker containers using AWS Batch.
8. Amorphic S3 is used to store intermediate files for the state machine execution jobs.
9. Optionally, the last tool in the state machine execution workflow uploads the report to the Laboratory Information Management System (LIMS).

10.An additional step is added to run an AWS Glue workflow to convert the VCF to Apache Parquet, write the Parquet files to a data lake bucket in Amazon S3 and update the AWS Glue Data Catalog.

11A bioinformatic scientist works with the data in the Amazon S3 data lake using Amazon Athena via a Jupyter notebook, Amazon Athena console, AWS CLI, or an API. Jupyter notebooks can be launched from either Amazon SageMaker or AWS Glue. You can also use Amazon SageMaker to train machine learning models or do inference using data in your data lake.



Thanks!

Do you have any questions?