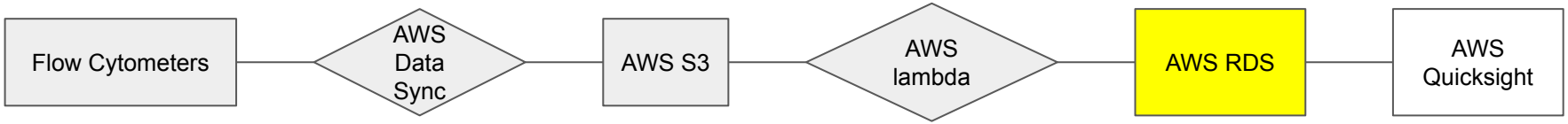
1. Database Design

I will also demonstrate a full extract transform load (ETL) operation within AWS to move the data into a database plus analytics on it after.



1. **AWS DataSync:** Allows for the connection of the laboratory cytometers into the AWS cloud environment. Can enable direct and secure sync between laboratory files generated by cytometer and AWS environment. This process automates file transfers between machine and cloud environment.
2. **AWS S3:** Cloud storage location where raw files generated from experiments are stored upon completion. Files organized by experiment type and Experiment/project ID and automatically loaded from AWS IOT core in step 1.
3. **AWS Lambda:** this is a function that preprocesses the raw data in S3 to format it for relational database in next step. If more complex preprocessing needs to occur on the data then AWS glue can be used.
4. **AWS RDS:** Relational database on MySQL engine
5. **AWS Quicksight:** Analytics dashboard able to display PCA, clustering, and analytical reports in automated ETL process sample to answer generation.

Machine Log Table (Laboratory System level)

- Machine_ID (Primary Key): serial number of the Cytometer
- Location: Physical location of machine
- Machine Type: Manufacturer info
- Last Calibration Date: date when last calibrated/maintenance
- Error_log: last warning or error message

Subjects Table

- Subject ID (Primary Key)
- ProjectID (Foreign key)
- Age
- Sex
- Condition
- Treatment
- Reponse

Projects Table

- ProjectID (Primary Key): A unique ID for each project.
- Machine_ID (Foreign Key): serial number of Cytometer
- Status: Current status of the project (for example Active, Completed, Paused).
- Samples: number samples run
- ExperimentmType: Type or category of the experiment.
- Lab Technician Name

Samples Table

- SampleID (Primary Key): A unique identifier for each sample.
- SubjectID (Foreign Key): Links to the Subjects Table.
- SampleType: Type of sample
- Time_from_treatment_start: time from when the patient first started treatment
- B_cell: B cell count
- Cd8_t_cell: Cd8 T cell count
- Cd4_t_cell: cd4_t_cell count
- Nk_cell: NK cell count
- Monocyte: monocyte count

2. What would be some advantages in capturing this information in a database?3. Based on the schema you provide in (1), please write a query to summarize the number of subjects available for each condition.

Advantages:

1. **Security and compliance**
    a. **When handling de identified PHI and security concerns it is most proper to store data in cloud environment with end to end encryption and security. Cloud platforms provide this under the shared data responsibility. For example, a Business Associate Agreement with AWS will allow the transfer and storage of PHI using their platform.**
2. **Capturing record of historical data**
    a. **For future cases depending on wording for customer contracts Teiko could amass a powerful database of experiments regarding flow cytometry. The dataset could be extremely valuable as it grows allow Teiko to provide more powerful analytics to customers**
    b. **The database also provides structured and a standard access with common SQL protocol**

```
SELECT Condition, Count(SubjectID) as NumberSubjects

FROM Subjects

GROUP BY Condition;
```

4. Please write a query that returns all melanoma PBMC samples at baseline (time_from_treatment_start is 0) from patients who have treatment tr1.

```
SELECT Subjects.SubjectID, Subjects.Age, Subjects.Sex, Subjects.Condition, Subjects.Treatment,
Subjects.Response, Samples.SampleID, Samples.SampleType, Samples.Time_from_treatment_start,
Samples.B_cell, Samples.Cd8_t_cell, Samples.Cd4_t_cell, Samples.Nk_cell, Samples.Monocyte
FROM Subjects
INNER JOIN Samples ON Subject.SubjectID = Samples.SubjectID
WHERE Subject.Condition = 'melanoma'
  AND Subject.Treatment = 'tr1'
  AND Samples.SampleType=’PBMC’
  AND Samples.Time_from_treatment_start = 0;
```

5. Please write queries to provide these following further breakdowns for the samples in (4):

a. How many samples from each project

```
SELECT COUNT(Samples.SampleID)
FROM Subjects
INNER JOIN Samples ON Subjects.SubjectID = Samples.SubjectID
WHERE Subjects.Condition = 'melanoma'
  AND Subjects.Treatment = 'tr1'
  AND Samples.SampleType='PBMC'
  AND Samples.Time_from_treatment_start = 0
GROUP BY Subjects.ProjectID;
```

b. How many responders/non-responders

```
SELECT
    COUNT(Distinct(Subjects.Response) )AS Responders
FROM
    Subjects
INNER JOIN
    Samples ON Subjects.SubjectID = Samples.SubjectID
WHERE  Subject.Condition=’Melanoma’
   AND Subject.Treatment = 'tr1'
   AND Samples.SampleType=’PBMC’
   AND Samples.Time_from_treatment_start = 0;
```

c. How many males, females

```
SELECT
    COUNT(Distinct(Subjects.Sex) )
FROM
    Subjects
INNER JOIN
    Samples ON Subjects.SubjectID = Samples.SubjectID
WHERE  Subject.Condition=’Melanoma’
   AND Subject.Treatment = 'tr1'
   AND Samples.SampleType=’PBMC’
   AND Samples.Time_from_treatment_start = 0;
```