**Inception Planning Sessions – Pre Sprint 1 Week 8**

Things to go over with the students:

ETL vs ELT – They need to use ETL –

Pandas or not – preferably not

**MVP** - Breakdown of ETL steps required in the projects

The teams need to produce a system which will do these things in turn:

**Extract:**

Read in a Transaction (CSV)

**Transform:**

Remove PII

Normalise to 3NF (transaction, basket, location)

**Load:**

Load into DB (RedShift)

Visualise (Grafana):

Volume (Product / Location)

Revenue (Product / Location)

Top Sellers (Product / Location)

Infrastructure (CloudWatch), e.g. lambda invocations, processing time, lambda errors.

**Super Cafe Product Client Requirements**

Based on the sales data from each branch that is provided in RAW CSV files, for each row in each CSV file you must.

Remove the Customer Name and the CARD Number - you should not store those anywhere.

Keep a Cash/Card payment method

Reorganise the transaction data so you split out the individual product from each order.

e.g. "Regular Flavoured iced latte - Hazelnut - 2.75, Large Latte - 2.45" into separate data rows for reporting on, normalise this.

This is because they want reports/graphs/visualisations later, on things like:

How many coffees did we sell, of which type, each week?

Which store had the highest sales each week, or day?

What was the total value of Hazelnut Coffee sales each week, totalled for all stores?

What was the total value of Large Latte sales each day, totalled for all stores?

They also wish you to build visualisation of you Cloud Architecture performance and operations

A diagram of a software process

Description automatically generated

**Approach to Final Project - Things to think about**

1. Break down Requirements -

Whole Project - What the User Wants

Technically What's needed

Functions

Input - CSV - Dummy Data

Output - Data to Database - SQL Statements

Processing- Transformation Rules - Quality of Data

Data Storage Options

Program Data Types/ Data Structures

File Types

Database Type (SQL no SQL)

Database Structures (Tables)

Cloud Components

EC2

S3

Lambda

SQS

Redshift

Automation

Sell Scripts

CloudFormation

CI/CD Pipeline

Create your own picture of Solution (ETL Pipeline)

What issues could you face

How can you mitigate them

Possible/Alternate solution Solutions

**Approach** - *Some additional Thoughts*

Slice up the ETL Pipeline-

**Extract Phase**

What Code do you already have that might work.

What new code might you add?

Can you logically list what an extract looks like?

**Transform Phase**

What Code do you already have that might work

What new code might you add?

Can you logically list what a transform looks like?

**Load Phase**

What Code do you already have that might work

What new code might you add?

Can you logically list what an extract looks like?

**Some Questions to think about: -**

Could you build a mock-up of a simple solution

What testing might you use

What Infrastructure and tools could you use

What libraries might you need