Pacific Infra Group 2: Divya Dhar, Aikya Shah

Business Info: Like Salesforce - provides a SaaS platform to other companies. Companies can sign up, provide licenses to a certain number of their staff, as well as buy additional features/support

5 Pipelines

1. Profit = subscription costs paid - expenses on all accounts
    - Unit level profit: profit / # of subscribers (+ info about cost per account)
2. Growth:
    - Increase in number of accounts per month, increase in size of accounts that are about to renew (increase in $ increased frrom upgrade in account)
3. Engagement
    - How many users using technology in company
    - How many hours per day are all users on account spending time on account

- Aggregate pipeline to Executives/CFO (ultimately presented to investors)
    - Weekly
- Aggregate pipeline to Experiment team
    - Data science team uses unit level/daily level data to conduct experiments on AB testing features being rolled out to different accounts
    - Weekly preferred, monthly - to drive direction of product team

The pipelines that will affect investor's are the Profit, Growth, Engagement and Aggregate Pipeline to Investors - so the next few pages will be runbooks for each pipeline

Runbooks

1. Pipeline Name: Profit
2. Types of data:
    a. Revenue from accounts
    b. what is spent on assets, other services according to Ops team
    c. Aggregated salaries by team
3. Owners: Finance Team/Risk Team
    a. Secondary Owner: Data Engineering
4. Common Issues:
    a. Numbers don't align with numbers on accounts/filings - these numbers need to be verified by an accountant if so
5. SLA's:
    a. Numbers will be reviewed once a month by account team
6. Oncall schedule
    a. Monitored by BI in profit team, and folks rotate watching pipeline on weekly basis. If something breaks, it needs to be fixed

1. Pipeline Name: Growth
2. Types of data: Changes made to the account type,
    a. # of users with license increased
    b. Account stopped subscribing
    c. Account continued subscription for the next calendar year
3. Owners: Accounts Team
    a. Secondary Owner: Data Engineer Taem
4. Common Issues:
    a. Time series dataset - so the current status of an account is missing since AE team forgot to
        i. A clue that it's missing is a previous step that is required in is missing (ex: only changes A, C, when step B is required to change to C)
5. SLA's:
    a. Data will contain latest account statuses by end of week
6. Oncall schedule
    a. No on call if pipeline fails, but pipeline will be debug by team during working hours

1. Pipeline Name: Engagement
2. Owners: Software Frontend Team
    a. Secondary Owner: Data Engineer Team
3. Engagement metrics come from clicks from all users using platforms in different teams

    a. Sometimes data associated with click will arrive to kafka queue extremely late - much after the data has already been aggregated for a downstream pipeline
    b. If kafka goes down, all user clicks from website will not be sent to kafka, therefore not sent to the downstream metrics
    c. Sometimes the same event will come through the pipeline multiple times - data must be de-duplicated
4. SLA's:
    a. Data will arrive within 48hrs - if latest timestamp > the current timestamp - 48 hrs, then the SLA is not met
    b. Issues will be fixed within 1 week
5. Oncall schedule:
    b. One person on DE team owns pipeline each week - there is a contact on SWE team for questions
    c. Next week - 30 min meeting to transfer onboarding to the next person

1. Pipeline Name: Aggregated data for executives and investors
2. Owners: Business Analytics team
    a. Secondary Owner: Data Engineer team
3. Common Issues
    a. Spark joins to join accounts to revenue, and engagement may fail - a lot of data is involved in the joins and there may be OOM issues
    b. Issues with stale data with previous pipelines - queue backfills periodically
    c. Missing data may cause issues with NA or divide by 0 errors
4. SLA's:
    a. Issues will be fixed by end of month, when reports are given to executives and investors
5. Oncall schedule:
    a. Around last week of month, DE's are monitoring that pipelines of the data from the month are running smoothly