

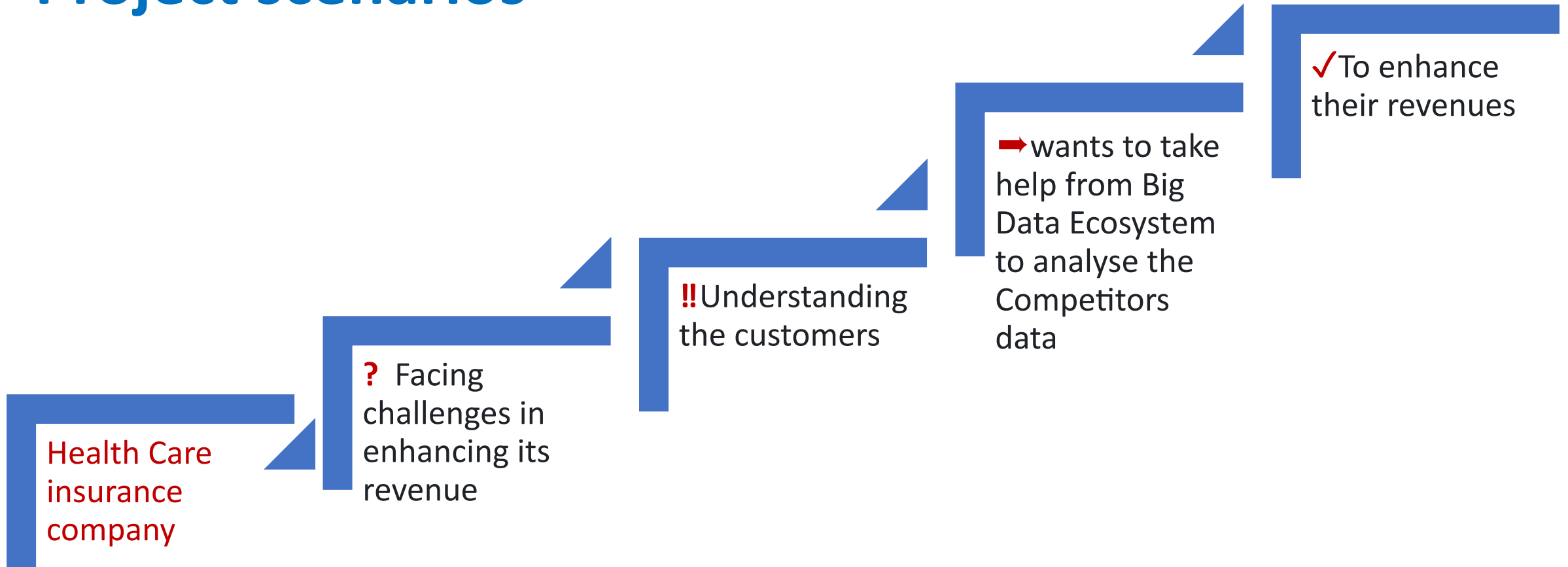
# Capstone Project

Pabitra Aryal

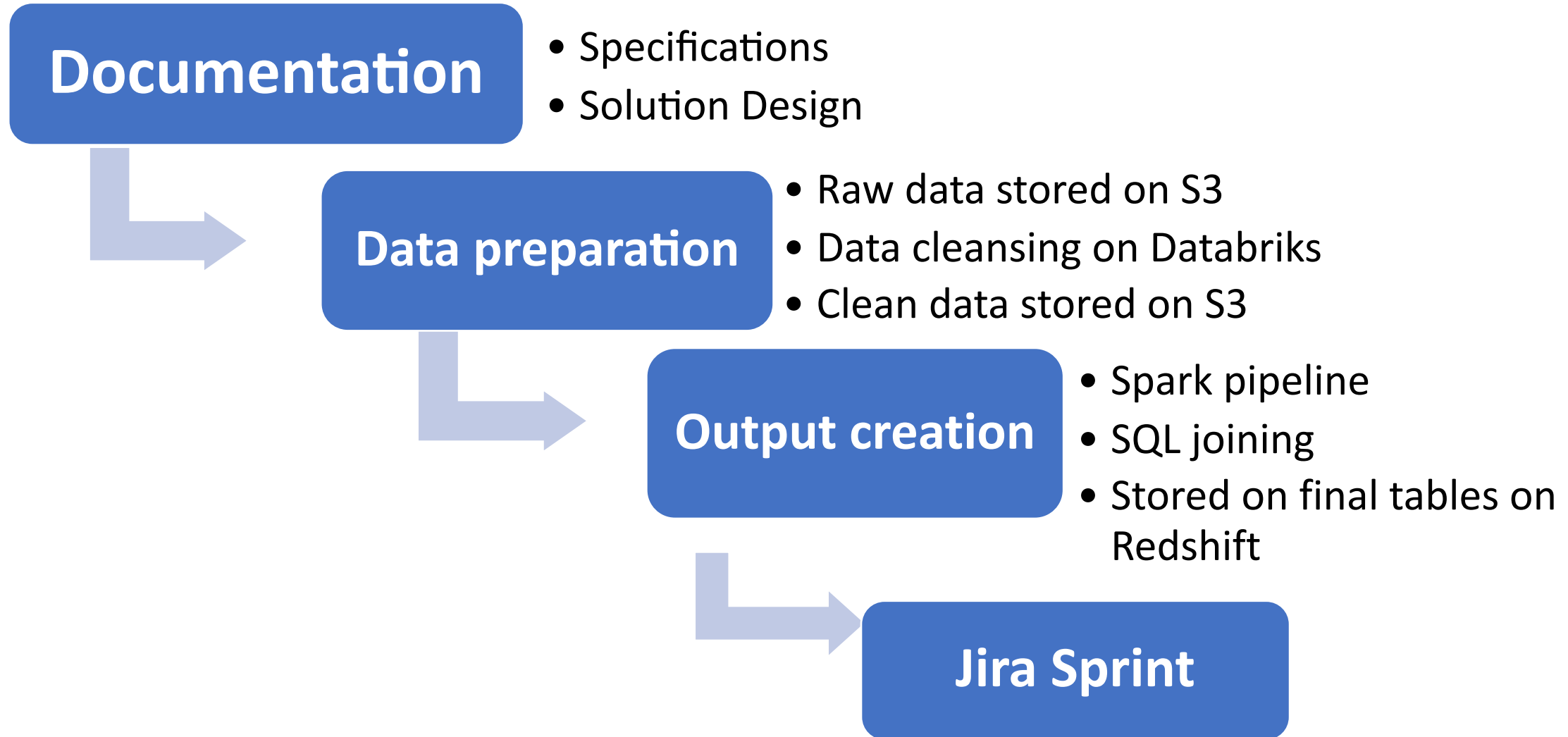
Takeo Bootcamp, BDE37

Date: 20 Dec 2023

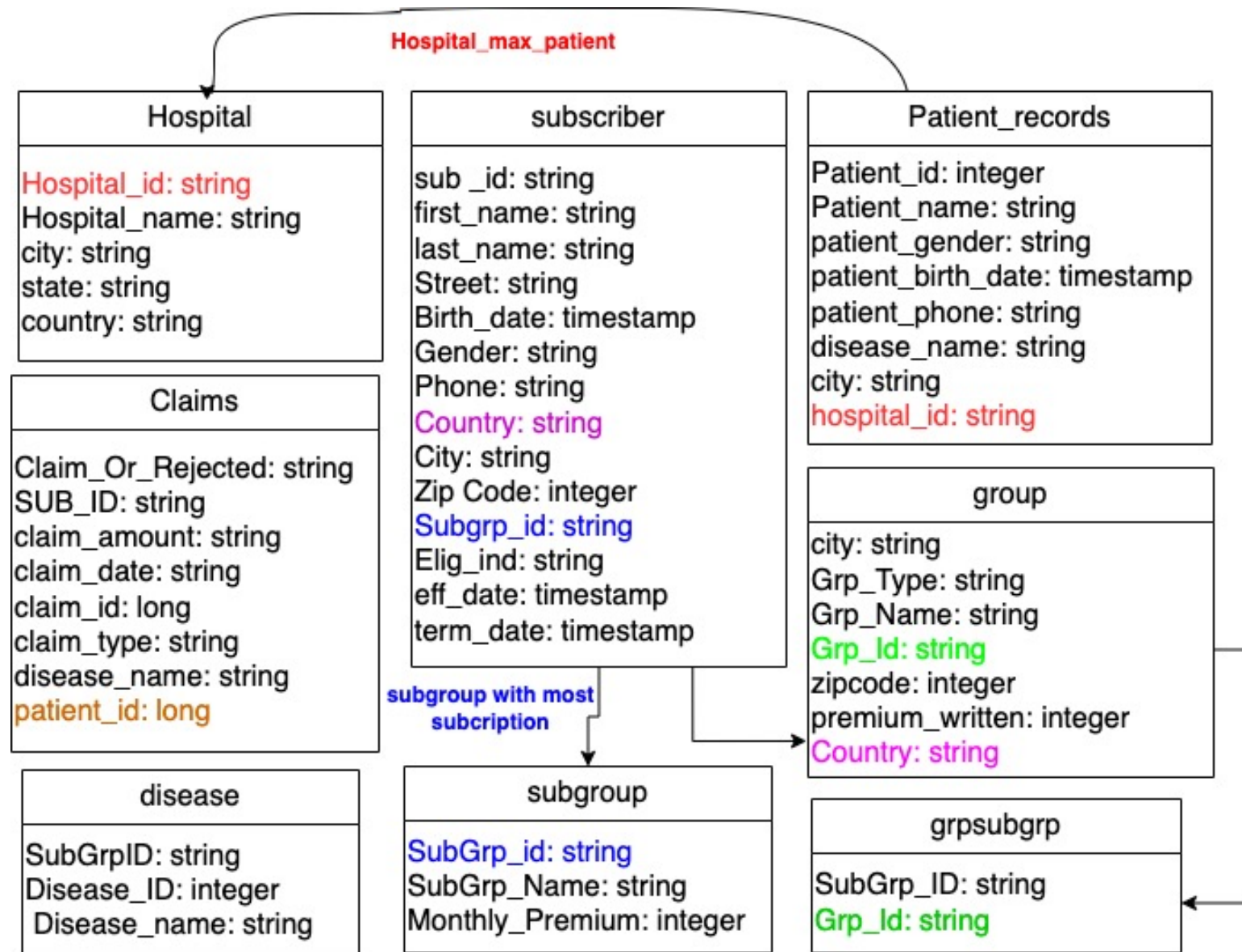
# Project scenarios



# Steps followed for the final results



# ER diagram



# Data preparation : *data cleansing*

```
#### Count string, timestamp, and integer null values
from pyspark.sql import functions as F
df.select([
    (
        F.count(F.when((F.isnan(c) | F.col(c).isNull()), c)) if t not in ("timestamp", "date")
        else F.count(F.when(F.col(c).isNull(), c))
    ).alias(c)
    for c, t in df.dtypes if c in df.columns
]).show()
```

Patient_id	Patient_name	patient_gender	patient_birth_date	patient_phone	disease_name	city	hospital_id
0	17	0	0	2	0	0	0

```
df = df.na.fill('NaN', subset=['Patient_name']) \
      .na.fill('NaN', subset=['patient_phone'])
df.show()
```

Patient_id	Patient_name	patient_gender	patient_birth_date	patient_phone	disease_name	city	hospital_id
187158	Harbir	Female	1924-06-30	+91 0112009318	Galactosemia	Rourkela	H1001
112766	Brahmdev	Female	1948-12-20	+91 1727749552	Bladder cancer	Tiruvottiyur	H1016

Amazon S3

Buckets

Access Grants [New](#)

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight 7

Amazon S3 > Buckets > bootcampapbitra > capstone/ > Input/

Input/

Objects

Properties

Objects (8) Info

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	claims/	Folder	-	-	-
<input type="checkbox"/>	disease/	Folder	-	-	-
<input type="checkbox"/>	group/	Folder	-	-	-
<input type="checkbox"/>	groupsub/	Folder	-	-	-

https://s3.console.aws.amazon.com/s3/#

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie p

Amazon S3

Buckets

Access Grants [New](#)

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight 7

Amazon S3 > Buckets > bootcampapbitra > capstone/ > cleaned\_input/

cleaned\_input/

Objects

Properties

Objects (8) Info

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	claims/	Folder	-	-	-
<input type="checkbox"/>	disease/	Folder	-	-	-
<input type="checkbox"/>	group/	Folder	-	-	-
<input type="checkbox"/>	groupsub/	Folder	-	-	-

5

# Use case solution



## Use\_Case\_solutions (Python)

[Import Notebook](#)

```
df.createOrReplaceTempView("patient_records")
df3=spark.sql("with T as (select count(c.claim_id) cnt, p.city from claims as c inner join patient_records as p on c.patient_id=p.Patient_id group by p.city) select city, cnt as total_count from T where cnt=(select max(cnt) from T)")
df3.show()
```

city	total_count
Mysore	2
Amravati	2
Kamarhati	2
Jabalpur	2
Bihar Sharif	2
Ghaziabad	2
Morbi	2
Karimnagar	2

```
##### Use case 7 ##### uploading to Redshift #####
df3.write.format("redshift").option("url","jdbc:redshift://default-workgroup.985881216142.us-east-1.redshift-serverless.amazonaws.com:5439/dev").\
    option("dbtable", "project_output.most_claimed_city").\
    option("aws_iam_role", "*****").\
    option("driver","com.amazon.redshift.jdbc42.Driver").\
    option("tempdir", "s3a://pabitrabucket/tmpdir/").\
    option("user", "*****").\
    option("password", "*****").save()
```

## Output creation: *final tables on Redshift*

aws

Services

Search

[Option+S]

N. Virginia

Admin\_IAM @ 9858-8121-6142

Editor

Queries

Notebooks

Charts

History

Scheduled queries

Redshift query editor v2

CreateLoad data

Filter resources

pg\_auto\_copy

project\_output

Tables13

cashless\_charges\_greater\_equal\_50th

female\_ageover40\_kneesurgery

group\_avg\_premium\_pay

group\_with\_max\_subgroup

groupof\_subscribe\_privt\_govt

hospital\_max\_patients

max\_no\_claims

most\_claimed\_city

most\_profitable\_group

patient\_under18\_cancer

subcr\_less\_30\_any\_subgrp

subgroup\_subscribe\_mosttimes

total\_rejected\_claims

Untitled 9 xUntitled 8 xUntitled 1 xUntitled 7 xUntitled

RunLimit 100ExplainIsolated session

Serverless: de...dev

Schedule

1SELECT

2\*

3FROM

4project\_output.most\_profitable\_group;

Result 1 (3)

ExportChart

	grp_name	avg_premium_rs
	IndiaFirst Life Insurance ...	99000
	Cholamandalam MS Gen...	99000
	Raheja QBE General Insu...	99000

Elapsed time: 215 msTotal rows: 3

CloudShell

Feedback

© 2023, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCookie preferences

# Jira Sprint

Jira Software

Your work ▾

Projects ▾

Filters ▾

More ▾

Create

✦ Upgrade

Q

Search

?

P

capstone\_project1

Software project

PLANNING

Timeline

Backlog

Board

Goals

+ Add view

DEVELOPMENT

Code

Project pages

Add shortcut

Project settings

You're in a team-managed project

Learn more

Projects / capstone\_project1

Backlog

Q

P

Invite

Epic ▾

Type ▾

↗

⋮

📄

📈

⚙️

▼ CP1 Sprint 1 13 Dec – 20 Dec (10 issues)

000 Complete sprint ⋮

Complete the specification, and solution design documents for the project.

📄 CP1-1 Documentation

DONE ▾

-

P

✓ CP1-2 Specification document

DONE ▾

-

P

✓ CP1-3 Design Document

DONE ▾

-

👤

✓ CP1-11 storing\_clean\_data\_S3

DONE ▾

-

👤

📄 CP1-4 Data\_preparation

DONE ▾

-

P

✓ CP1-5 Upload data on AWS S3

DONE ▾

-

👤

✓ CP1-6 Data cleaning

DONE ▾

-

👤

📄 CP1-8 Result creation on AWS Redshift

DONE ▾

-

P

✓ CP1-9 Spark pipeline

DONE ▾

-

👤

✓ CP1-10 Upload results on Redshift

DONE ▾

-

👤

💡 Quickstart

✕

8



Thank you for listening!