**High Level Specification:**

1) Data must be transferred between source data systems and the Redshift data warehouse.

2) There must be flexibility to add multiple source systems (e.g. MongoDB, MySQL, etc.) simply

3) There must be no data loss through the ingestion processing - appropriate logging & checksums should be carried out to ensure that this is the case

4) In the case of pipeline failure, we should have notification to the team to ensure that the issues can be resolved & data can be re-processed

5) Duplicate rows of data should be handled & removed during processing to ensure we have no data duplication

6) Data must be validated (e.g. domain constraint tests) & outliers / perceived issues should be logged

7) Data must be ingested into Redshift within 30 minutes

8) The data must be loaded into a secure schema, with access only provisioned to teams X, Y and Z.

**QA Requirements:**

1) We should unit test each component. We can write test cases to ensure we are testing functionality & data validity against a controlled test set of data

2) We should end-to-end test the process of pulling data from MongoDB and loading it into Redshift, with a controlled set of test data that mimics the format & load of the production dataset

3) We should test performance of the pipelines to ensure data is available within the required timeframe.

4) We should simulate errors to ensure that our error handling provisions are robust enough & to ensure that our logging & monitoring processes work as expected.

5) We should regression test with each deployment to ensure that existing functionality has not broken while deploying new functionality

6) We should test access controls to ensure access to data is restricted