Team Members:

1. **Mihir Thakkar (mt1104)**
2. **Utsav Patel (upp10)**

Project Deliverables:

1. ***A clear, specific, measurable technical goal you can realistically achieve by the end of the semester:***

   To analyze the following different map-reduce design patterns on the **metric of time** for the **sales product dataset**:

   1. Simple MapReduce design
   2. In-Mapper design
   3. Combiner based design

   Statistics to be computed using these patterns:

   i. **Item Count:** Returns the count of top K selling items in the dataset.
   ii. **Annual revenue generated by each item:** Return top K items in terms of annual revenue.
   iii. **Monthly Total Sales:** Returns the total sales in each month
   iv. **Item-wise monthly revenue**
   v. **Top K items** contributing to the monthly revenue for all months.

2. ***Existing prior work towards this goal:***

   Since this is a comparative analysis project, we do not have any prior work to refer to.

3. ***A description of how your project is novel, different, and/or builds on this prior work (there must be something novel about what you do, however minor or incremental; just state it precisely)***

   Our idea is to compare the different map reduce design patterns for a specific dataset to extract different aggregations from the dataset (refer to the statistics mentioned above). It is an analytical comparative approach to understand the effectiveness of different design patterns in solving the same problem.

4. ***A brief outline of the key technical idea or approach you intend to implement. Libraries we are going to use:***

   We intend to implement the different map reduce design patterns on a specific dataset to solve the problem of extracting different aggregate information.

Libraries:

A. mrjob

B. Numpy

C. Pandas

5. ***Metric and Method to be used:***

We will register the time taken to complete the task using different map-reduce design patterns.

6. ***The risks that may prevent you from achieving your stated goal:***

1. Due to complex internal implementation the mrjob library, we are not sure about the time it may take to run the aggregation processes.
2. Given the limited availability of ilab resources and a lot of users of these resources, we might not be able to access ilab whenever needed.