# Geo Hash-Query Layer

## Motto:

Leverage the knowledge of previous partitioning and efficiently query on subset of data:

- Identify the type of Query
- Determine the areas of interest from the overall world
- Figure out partitions that contain our interested region
- Slice out the partitions from the RDD
- Push the query on the final RDD based on the Query predicates

## Steps Explained:

Note: We have already completed, Stage 1 - executing the load aware Geohash partitioner on the data-set and immediately PERSISTING. This is the parent RDD we are working on.

> ***Warning:*** Failure to persist an RDD after it has been transformed with partitionBy() will cause subsequent uses of the RDD to repeat the partitioning of the data(reevaluation of the RDDs complete lineage). That would negate the advantage of partitionBy(), resulting in repeated partitioning and shuffling of data across the network, similar to what occurs without any specified partitioner.

**Stage-2:** Efficient Retrieval based on query

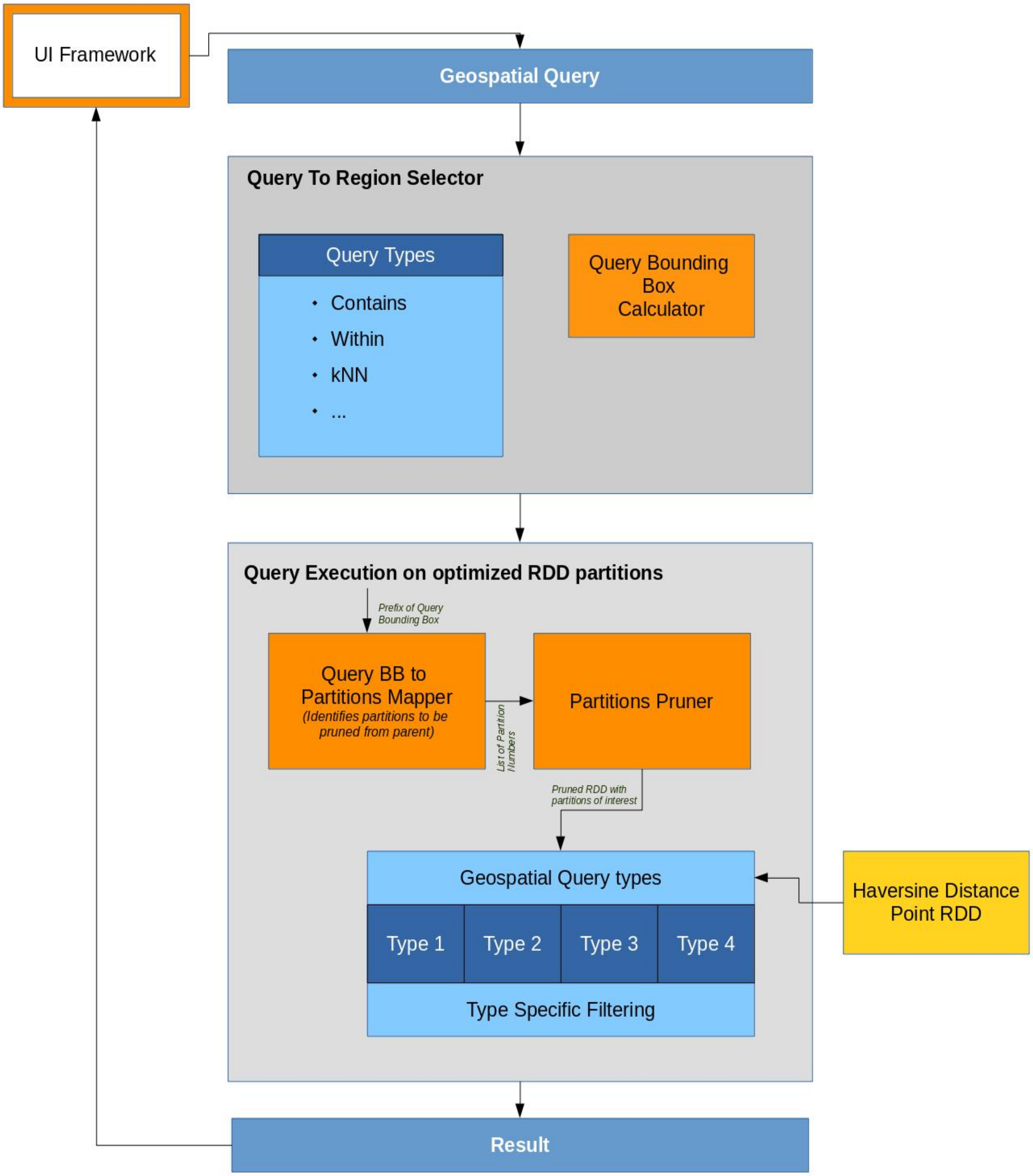
- Identify the type of Query
  - Find out the need operations
- Determine the areas of interest from the overall world
  - Determine the bounding box which covers all the regions/grid of interest. Based on

```
 Precision, Distance of Adjacent Cell in Meters

1, 5003530          6, 610
2, 625441           7, 118
3, 123264           8, 19
4, 19545            9, 3.71
5, 3803             10, 0.6
```

  - Fine-grained selection of candidate partitions would further improve performance
- Figure out partitions that contain our interested region
  - Once we have the bounding box encode string of the interested areas, compare it against the data structure used in Geohash partitioner(variable length key to partition map)
  - Select all the candidate partitions

- Slice out the partitions from the RDD
  - Use the list of candidate partitions to prune out the partitions from parent geohashed RDD.
- Push the query on the final RDD based on the Query predicates
  - Based on the type of query, filter out the records in RDD and present the result
  - [Convert to Data-frames if needed by the application for querying]


UI Framework
Geospatial Query
Query To Region Selector
Query Types
Contains
Within
kNN
...
Query Bounding Box Calculator
Query Execution on optimized RDD partitions
Prefix of Query Bounding Box
Query BB to Partitions Mapper
(Identifies partitions to be pruned from parent)
List of Partition Numbers
Partitions Pruner
Pruned RDD with partitions of interest
Geospatial Query types
Type 1
Type 2
Type 3
Type 4
Type Specific Filtering
Haversine Distance Point RDD
Result