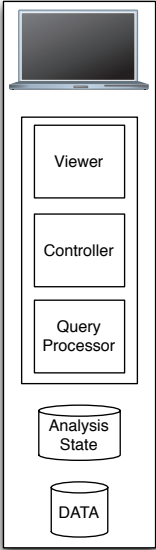# Interactive Analysis Design Sketch

Modes of Operation

Laptop Only
Application runs locally on user's laptop which also contains all the data


Laptop with Local Cluster
Similar to the laptop only case, but the application auto-discovers a local cluster which it can use to distribute the work of processing queries. The application replicates part of the data to each node of the cluster and then sends the query to be processed. Intermediate results can also be cached on the cluster nodes for use by future queries.
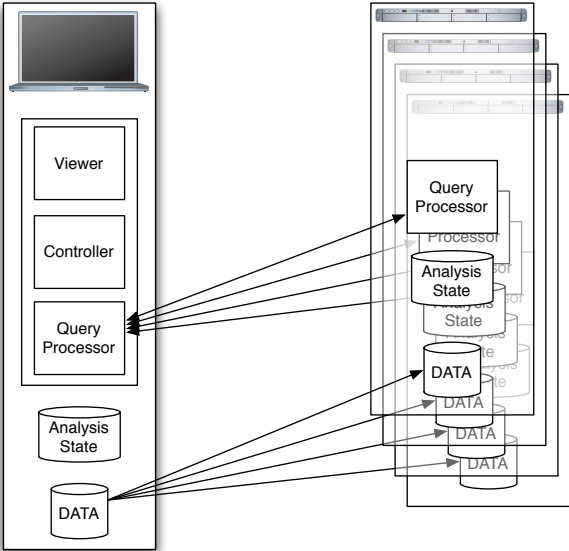
Interactive Analysis Design Sketch

Laptop using a Data Server and a Local Cluster
In this configuration, the laptop does not contain the data for the query, instead it is stored in a data server on the local cluster.
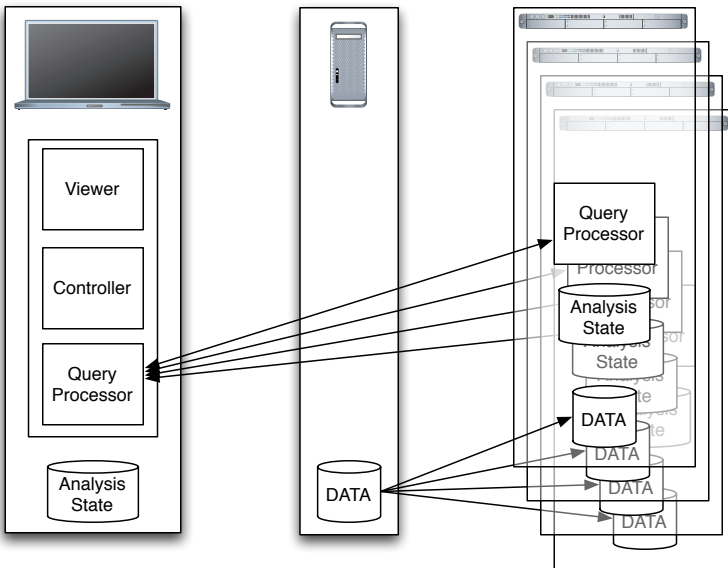
Application using a Web Service
Here the laptop does not need to be on the local network of the cluster and instead communicates via a web service. Analysis state could be stored on the Web Service for extended periods of time to allow for collaborative analysis.
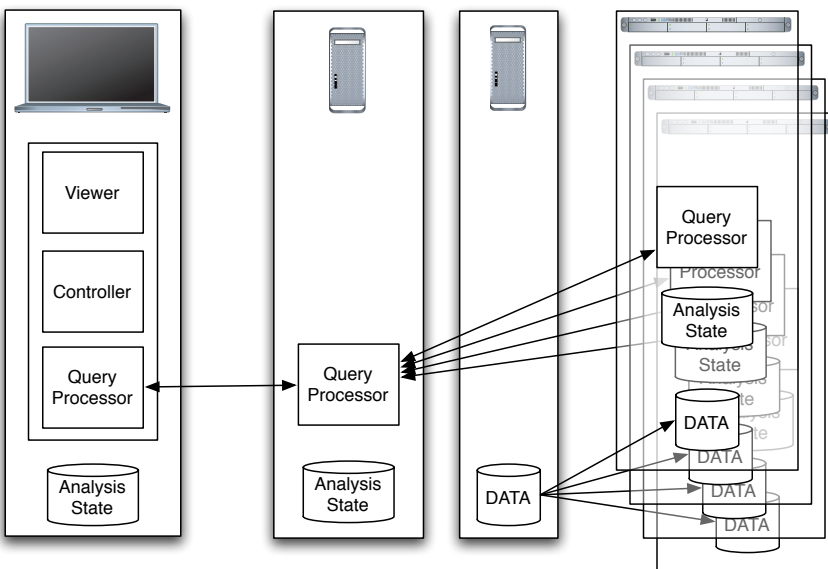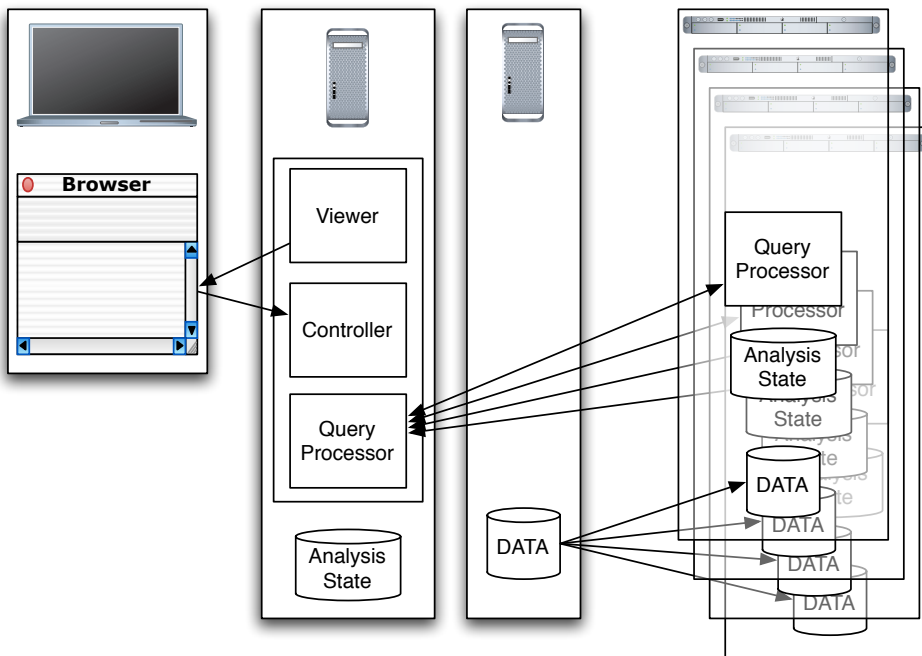

Web Browser using a Web Service
In this case, minimal processing is done on the laptop with the majority of the work being done remotely. This mode offers less interactive control but minimizes user's setup. [Think about the difference between Google maps in a browser and the Google Earth

Interactive Analysis Design Sketch
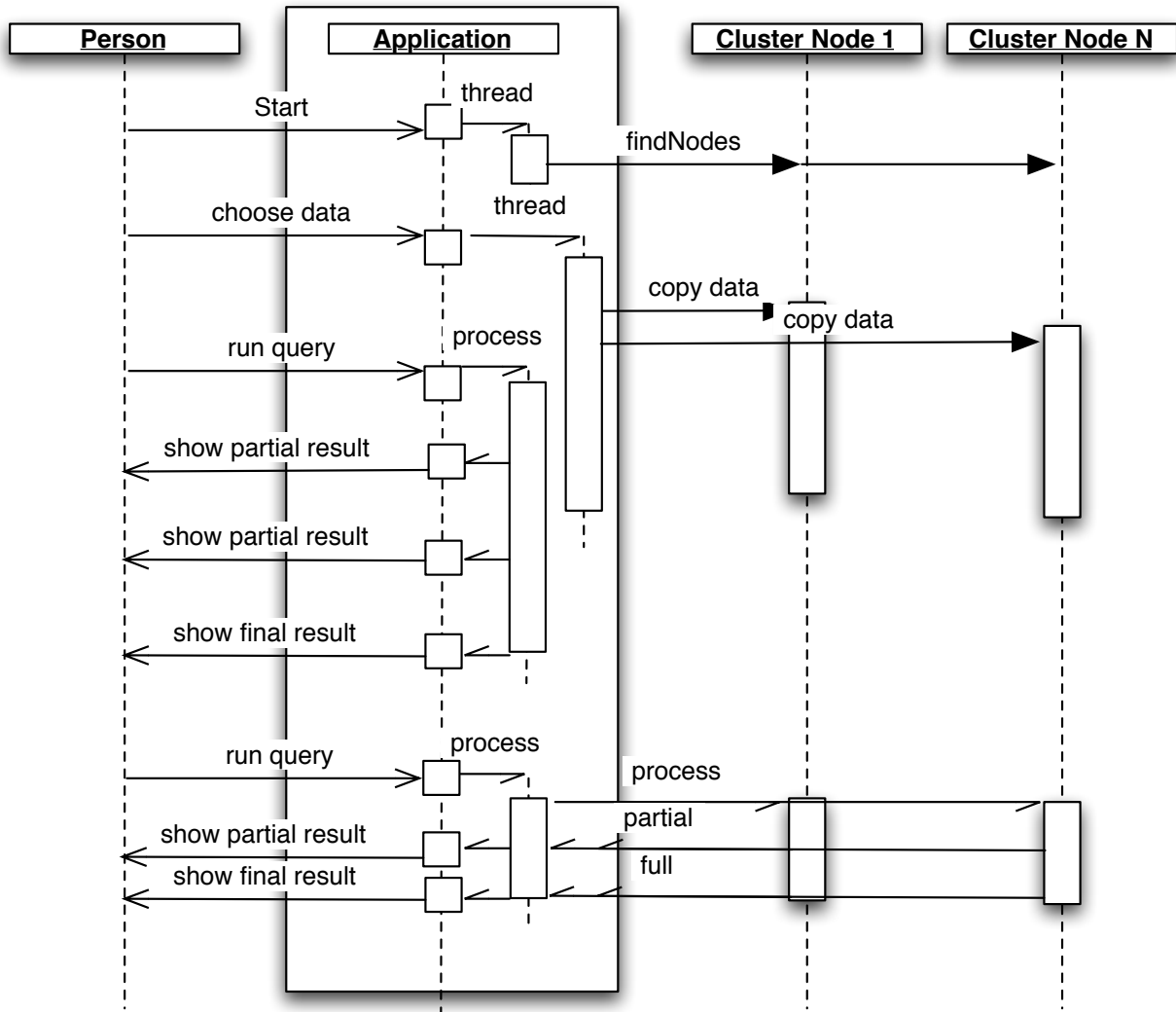
application]

System's Response to User Interactions

Simple Session Overview

Our goal is to have the system be as responsive to the user as we can possibly be. As part of that we should give a preview of the query results by showing partial results of a query the user has requested of the data. Not only does this allow users to know that the system is running but it also gives fast feedback about whether the query really is what the user wants to request of the data.

Below is a timing diagram of a hypothetical session. When the application first starts up it looks for the availability of a local cluster to help with the data processing. Once found it begins to replicate the data from the laptop to the cluster nodes. During the replication process, the user can still make queries, which will only be handled by the resources local to the laptop. Once the replication has finished, any further queries will be distributed to the cluster. This allows the user to be able to work continuously without having to wait for the system to fully set itself up.

Interactive Analysis Design Sketch

Person	Application	Cluster Node 1	Cluster Node N

Start
thread
findNodes

choose data
thread
copy data
copy data

run query
process
copy data

show partial result

show partial result

show final result

run query
process
process
partial
show partial result
full
show final result

Preemptive Processing of Queries and Provenance Tracking
Computers are faster than humans so while a user is constructing a query, the system
could begin to do preliminary work based on the parts of the query known to it. There-
fore by the time the user has finished with the query and has requested that it be run the
system will have done lots of the ground work to make that query run as fast as possi-
ble. In addition, the system keeps track of provenance of the initial data and of the in-
termediate results of previous queries in order to determine if an intermediate result can
be used to satisfy the present query. Tracking of provenance also allows a user to ask
exactly how a particular result was derived.

Interactive Analysis Design Sketch

Person	Query Constructor	Query Processor	Cache Manager	Provenance Manager

Start

Add criteria → using criteria → criteria within known cache? → criteria within provenance?

Add criteria → using criteria → criteria within known cache? → criteria within provenance?

Run query → Run → provenance id

results already cached?

appropriate dataset cache to use

run ()

show final query ← cache