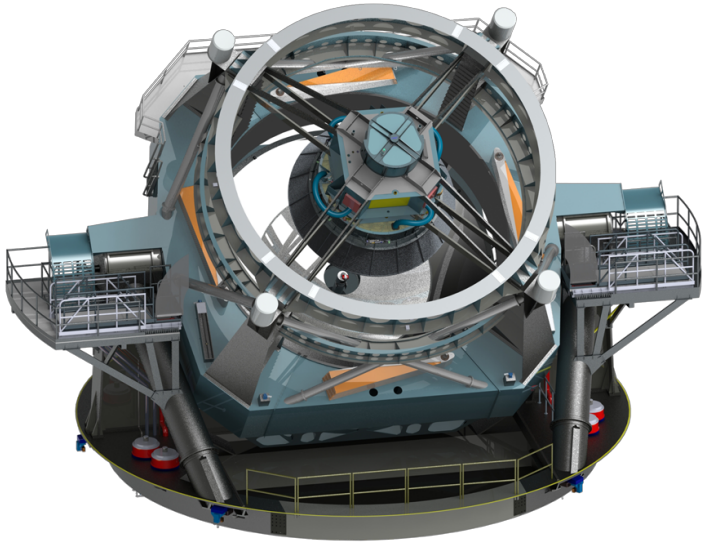
Large Synoptic Survey Telescope

Next-to-DB Questions



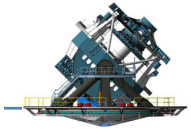
- **What is the data format?**

- Parquet is the de facto standard for modern work. Tested in Andy H's work, already used in Pipelines, and can be used as a replica copy for qserv

- Implies that data serving is something simplistic, think HTTP, object store.

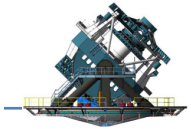
- **What is the computational library?**

- The leading contenders are dask and spark. Neither is an obvious winner over the other, and they are somewhat complimentary.
  - Spark: Robustness++, pythonicness--
  - Dask: Pythonic++, robustness--



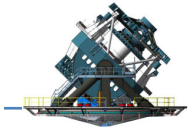
- **What is the user interface?**

- The notebook, either for dask or spark. This is the most natural in terms of the user experience, for sending code to the workers and for retrieving the result data.



What is the provisioning system?

- Dask supports calls to k8s to spin up worker pods, spark probably can do the same (starting tests off-project at UW). This may require a layer of intermediation for resource control or A&A.
- Dask and spark can also submit to a batch system
- Maintaining a cluster of workers shared across users (e.g. YARN) seems infeasible and not well suited to the way these libraries are designed.



- **What are the computational resources, quantitatively?**

- This is hard, and tightly coupled to the question of 10%.
- In current frameworks, very little is done to share or conserve resources between users.
- How do we handle concurrent usage? Do we let users have idle cores/storage in memory?