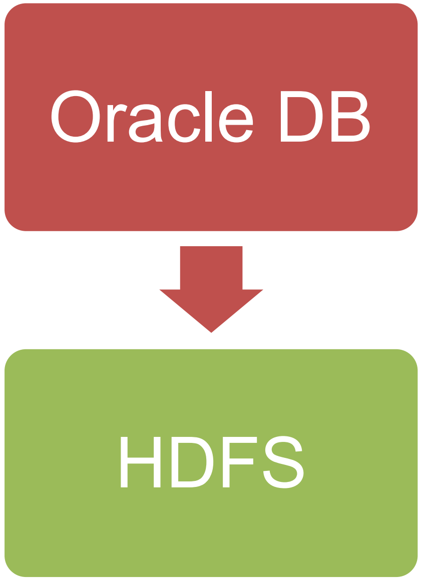
Extending PhEDEx monitoring

- We want to monitor space used by different datasets to plan distribution and cleaning

- Preserve replica history exporting daily snapshots to HDFS with Sqoop
- CERN IT *analytix* Hadoop cluster
  - 38 nodes, 2.8 PB
- Extends monitoring without impact on live service


Oracle DB

HDFS

27/09/16

Conclusion

- Enabled CMS dataset replica monitoring using analytics tools to aggregate and visualize data
- Efficient

  Can process 1 TB of input data for 1 year in 30 minutes
- Fully-covered

  We can afford to keep raw input data indefinitely
- Highly configurable

  Aggregations can be customized for specific analyses