



SmartData Fabric® on a MapR Data Lake Proof-of-Concept (POC)

February 2020



POC Goals

1. Establish SmartData Fabric® index and query processing layer on top of a MapR data lake containing customer data warehouse data
2. Demonstrate query functionality to support Tableau and other standard apps
3. Demonstrate performance and scalability on a large dataset
4. Demonstrate access security and data security to meet compliance, e.g., PCI and CCPA
5. Receive approval to implement solution as a production pilot

Background for Proof-of-Concept (POC)

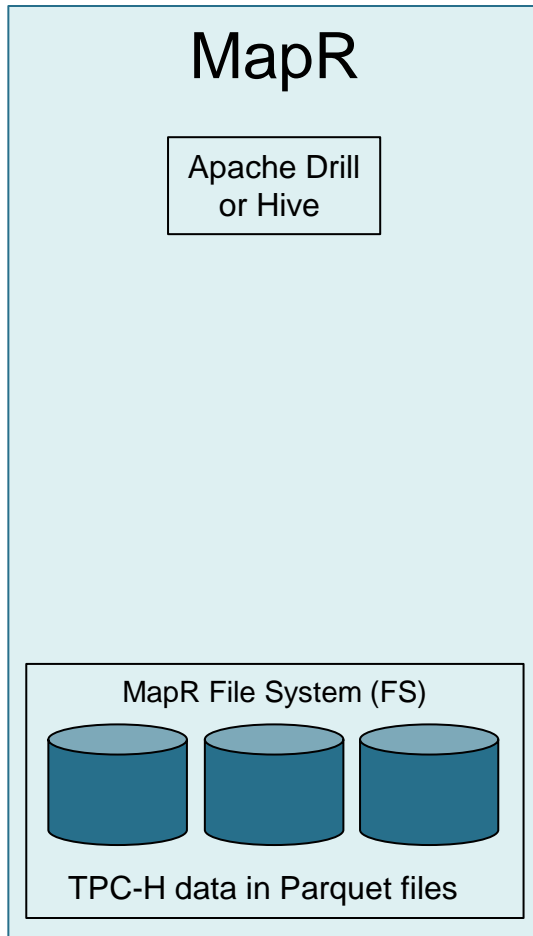
CONSTRAINTS

- Unable to use customer data
- Unable to install on customer systems

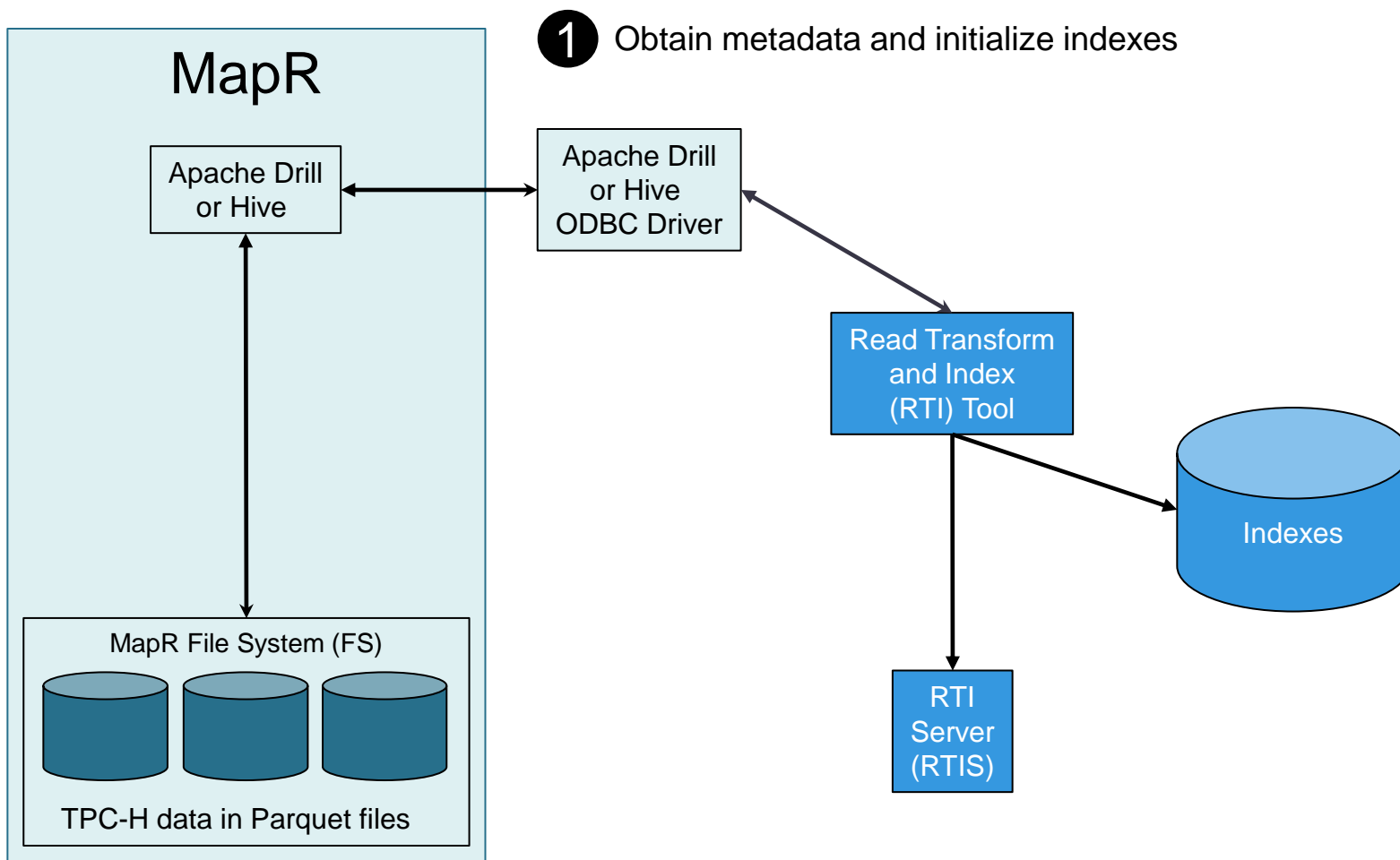
POC SOLUTION

- Build a POC external to customer environment
- Leverage commonly used test data from Transaction Processing Council Decision Support Benchmark (TPC-H, www.tpc.org/tpch/) consisting of multiple tables similar to a data warehouse
- Generate TPC-H data and store as commonly used Parquet files on a MapR File System (MapR-FS)
- Use a third-party ODBC driver from Apache Drill or Hive to read metadata from, and enable access to, MapR
- Index data from MapR and build indexed views
- Execute queries against MapR indexes and indexed views using Tableau in Live Connection mode

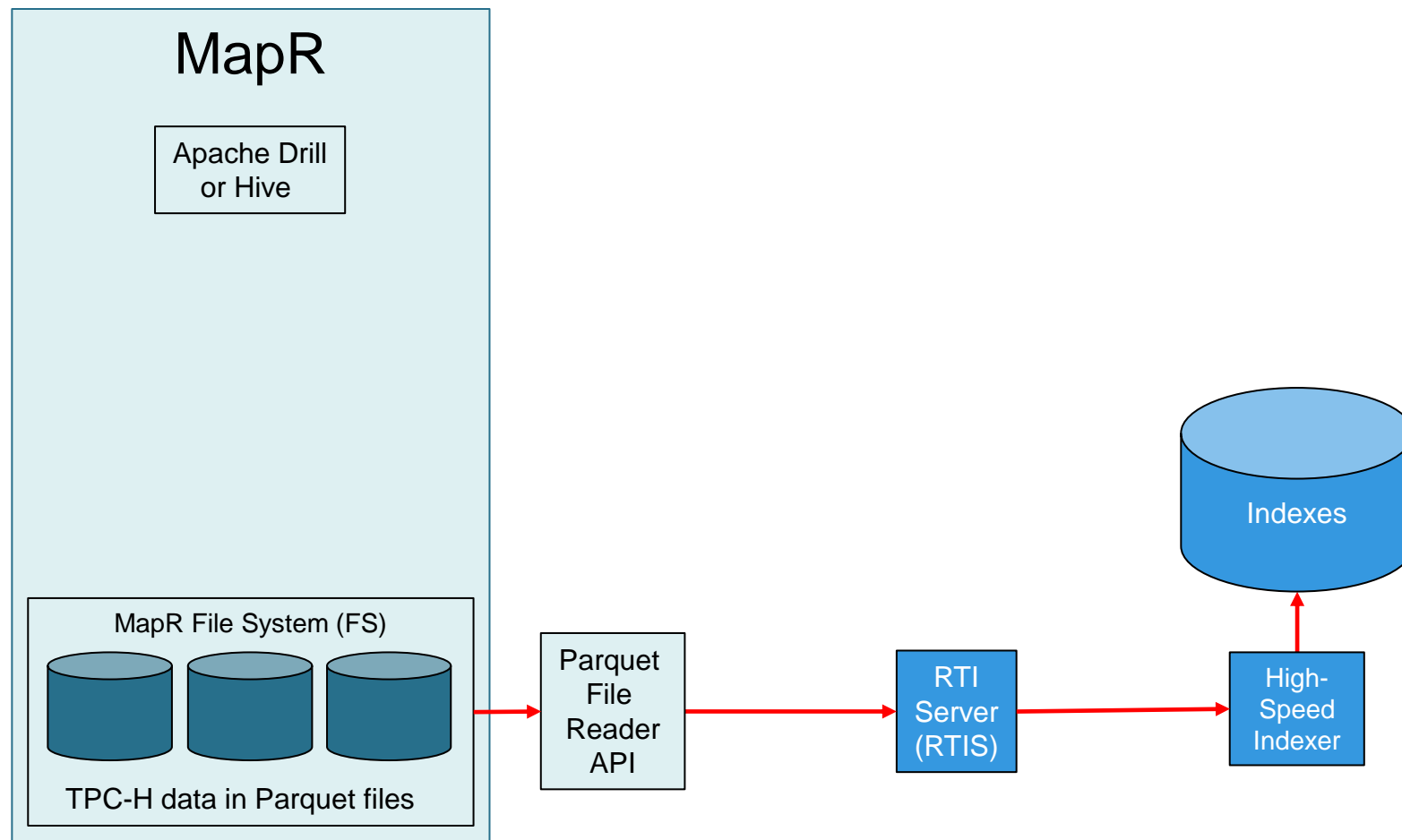
A MapR Data Lake



SmartData Fabric® on a MapR Data Lake – Step 1

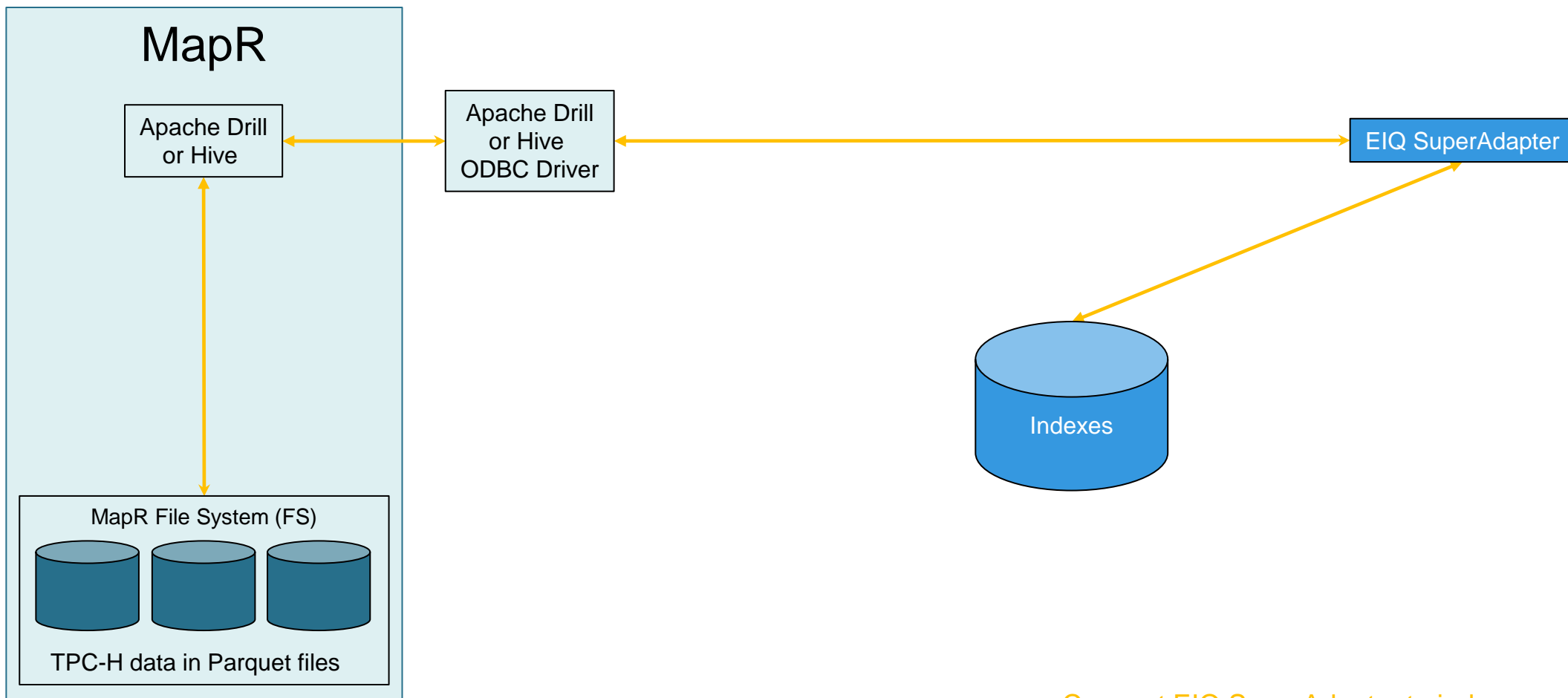


SmartData Fabric® on a MapR Data Lake – Step 2



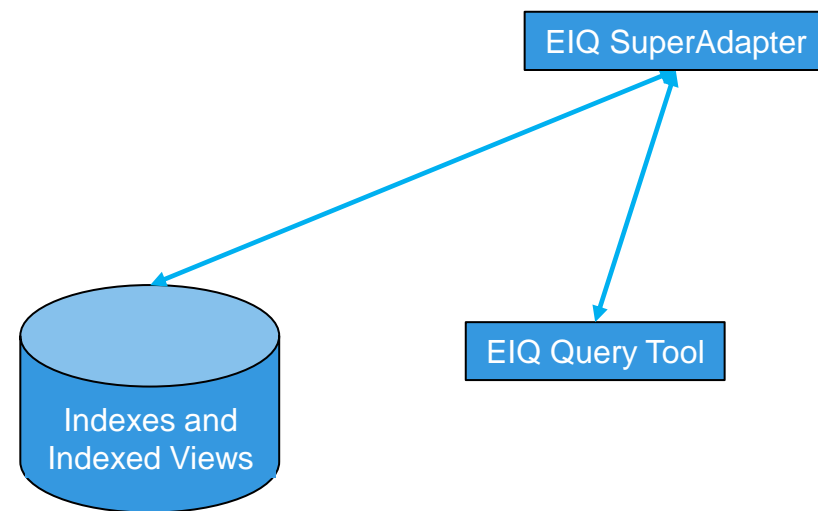
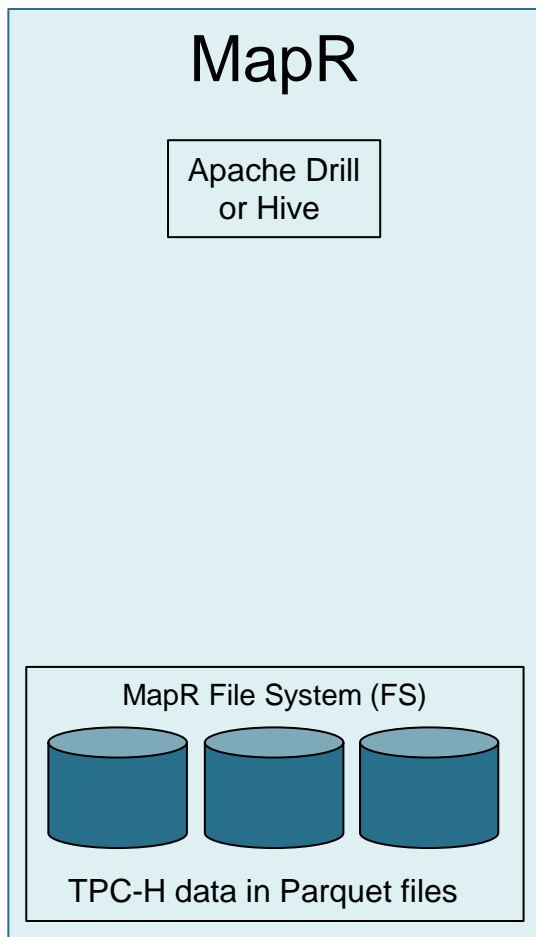
2 Build indexes

SmartData Fabric® on a MapR Data Lake – Step 3



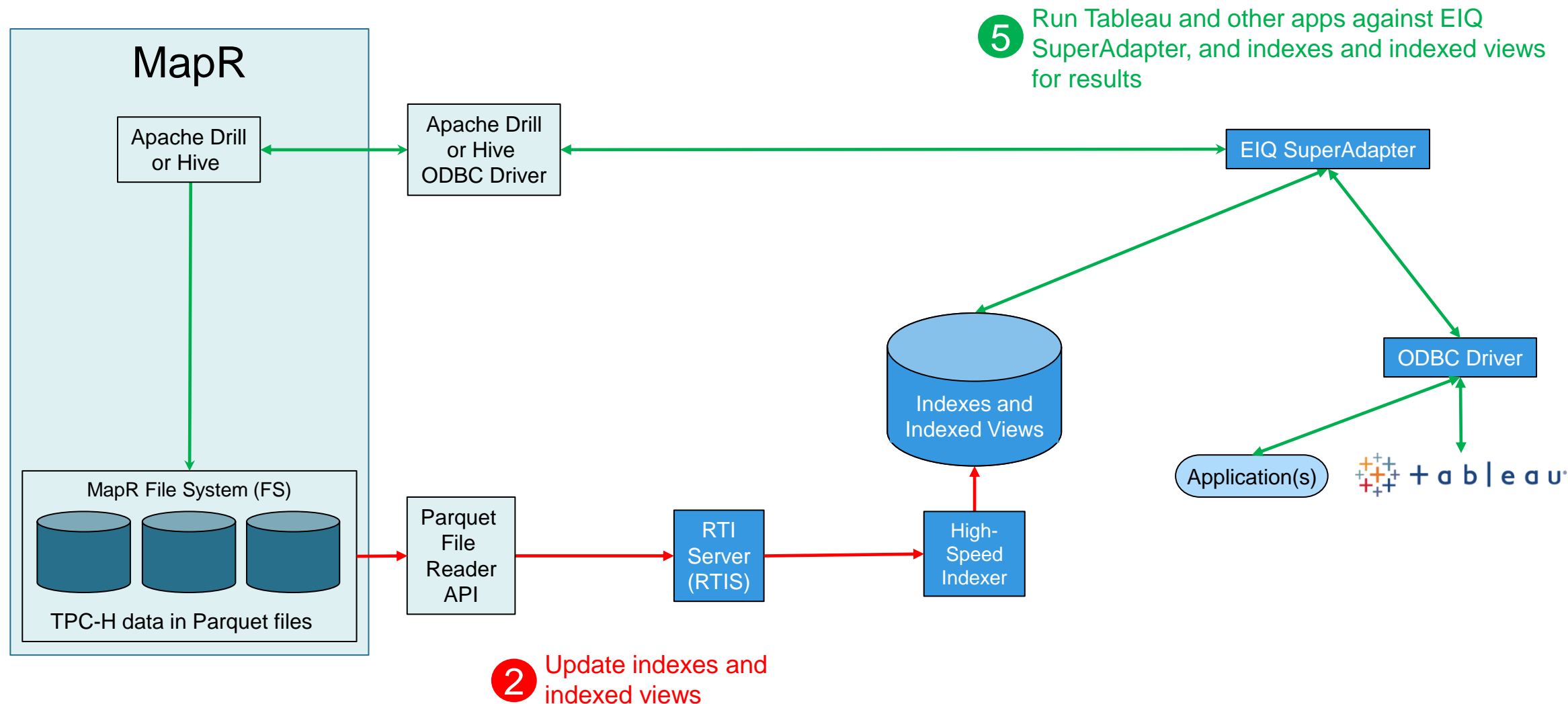
3 Connect EIQ SuperAdapter to indexes and map to Standard Data View (SDV)

SmartData Fabric® on a MapR Data Lake – Step 4

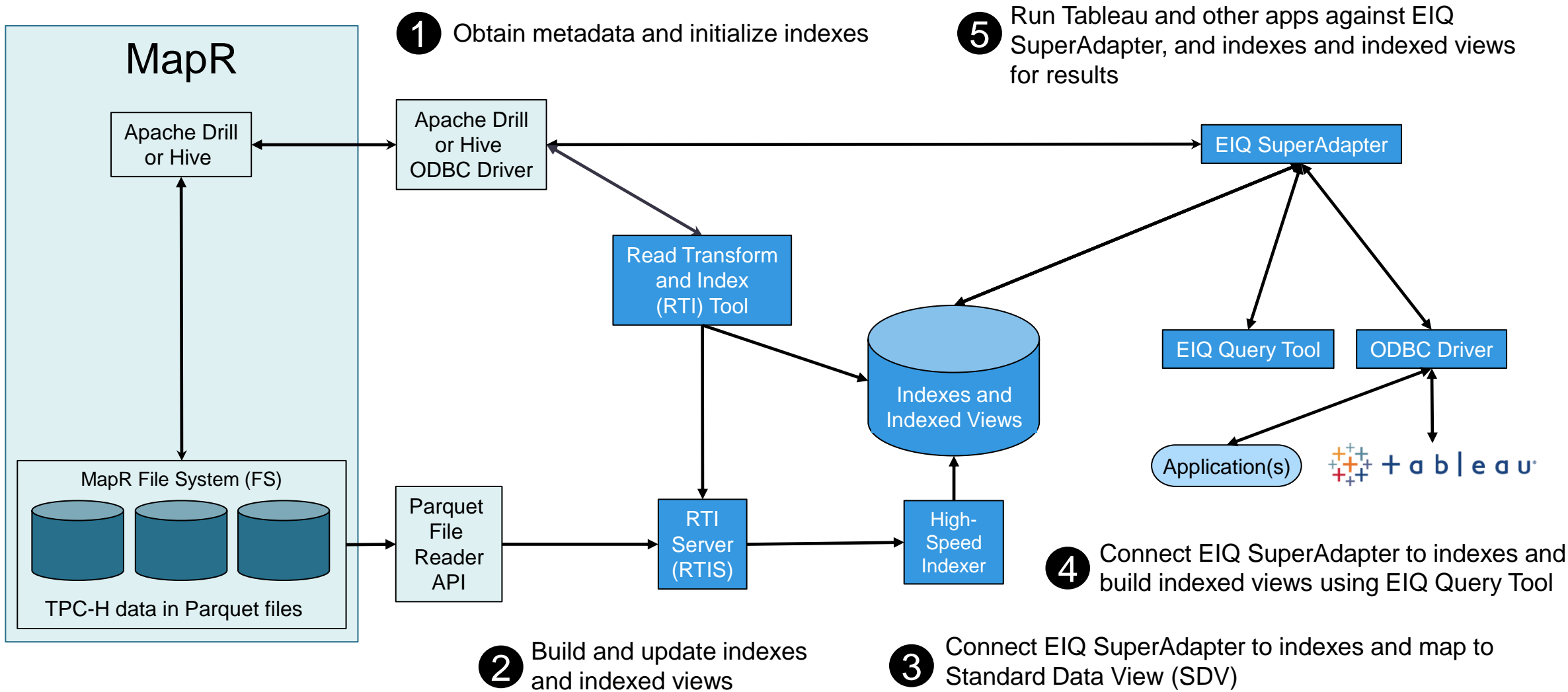


- 4 Connect EIQ SuperAdapter to indexes and build indexed views using EIQ Query Tool

SmartData Fabric® on a MapR Data Lake – Step 5



SmartData Fabric® on a MapR Data Lake - Overview



Future options

- Apache Drill and Hive ODBC drivers may or may not be the optimum means of access to Parquet files – currently testing
 - Depending on further tests, WhamTech may obtain or develop direct-to-Parquet files ODBC driver that can work seamlessly with SmartData Fabric® tools
- Use a means of Change Data Capture (CDC) or file-polling to obtain updates for maintaining indexes and indexed views in near real-time
- Add segments for production-level scaling



The End