Data Processing: Designing for Updates

Updated on: 1/5/2017

# Introduction

Many use cases around data life cycle management, such as loading of daily or hourly data, require that one can add individual partitions and manage them separately. This is best done with partitioning a table into individually addressable partitions. Each of these partitions have to then explicitly be added with ALTER TABLE ADD PARTITION and can be removed with ALTER TABLE DROP PARTITION. In addition, the query processor will perform partition elimination on supported predicates. Currently U-SQL only allows partitioning in conjunction with HASH and ROUND ROBIN distributions.

# Create a partitioned table for commits

* Submit the following script

USE DATABASE [ghanalytics];

@commits =

EXTRACT id int?,

sha string,

author\_id int?,

committer int?,

project\_id int?,

created DateTime?

FROM @"/TR24/GHData/Commits.csv"

USING Extractors.Csv();

@rs1to4 =

SELECT \*

FROM @commits

WHERE created < DateTime.Parse("1/1/2015");

CREATE TABLE Commits

( id int?,

sha string,

author\_id int?,

committer int?,

project\_id int?,

created DateTime?,

INDEX commits\_idx

CLUSTERED(id ASC)

PARTITIONED BY

BUCKETS(created)

HASH(id) INTO 5

);

DECLARE @cdate1 DateTime = new DateTime(2012, 1, 1, 00, 00, 00, 00, DateTimeKind.Utc);

DECLARE @cdate2 DateTime = new DateTime(2013, 1, 1, 00, 00, 00, 00, DateTimeKind.Utc);

DECLARE @cdate3 DateTime = new DateTime(2014, 1, 1, 00, 00, 00, 00, DateTimeKind.Utc);

DECLARE @cdate4 DateTime = new DateTime(2015, 1, 1, 00, 00, 00, 00, DateTimeKind.Utc);

ALTER TABLE Commits ADD PARTITION (@cdate1), PARTITION(@cdate2), PARTITION(@cdate3), PARTITION(@cdate4);

INSERT INTO Commits

ON INTEGRITY VIOLATION IGNORE

SELECT id,

sha,

author\_id,

committer,

project\_id,

created

FROM @rs1to4;

# Create and Load data into a new Partition

Add a fifth partition and statically load the data for the final partition into it.

Submit the following U-SQL Script.

USE DATABASE [ghanalytics];

DECLARE @cdate5 DateTime = new DateTime(2016, 1, 1, 00, 00, 00, 00, DateTimeKind.Utc);

@commits =

EXTRACT id int?,

sha string,

author\_id int?,

committer int?,

project\_id int?,

created DateTime?

FROM @"/TR24/GHData/Commits.csv"

USING Extractors.Csv();

@rs5 =

SELECT \*

FROM @commits

WHERE created >= DateTime.Parse("1/1/2015")

AND created < DateTime.Parse("1/1/2016");

ALTER TABLE Commits ADD PARTITION(@cdate5);

INSERT INTO Commits

PARTITION

(

@cdate5

)

SELECT id,

sha,

author\_id,

committer,

project\_id

// , created (do not specify, its explicit)

FROM @rs5;