MarketShare

Data Management on the Cloud

MarketShare

Defining the cloud

A **dynamically provisioned** commodity cluster of virtual machines with the following characteristics

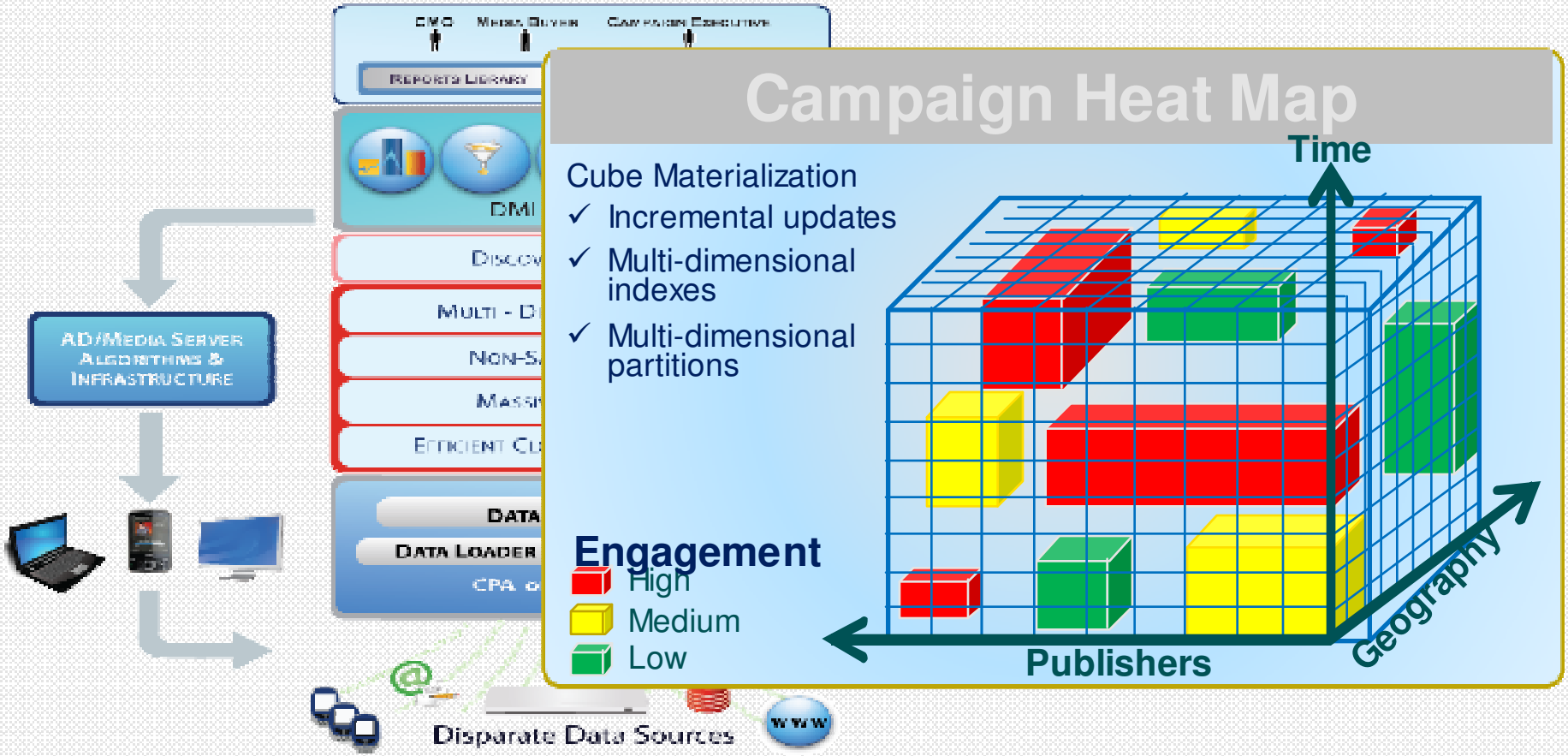
- **Infinite**
  - A large number of nodes can be commissioned in minutes
- **Taxi Meter**
  - Most services are billed at an hourly usage level

MarketShare

3 new problems for query processing

- **When** should resources be added to a data processing system?
  - Partition management for Low Cost on the cloud
- How many of these resources should be **permanent**?
  - Materialization with Intermittent Scalability
- **Where** should these resources be added in the stack?
  - Replication to Improve Query Performance

Transforming Data to Actionable Insights



MarketShare

The Schema

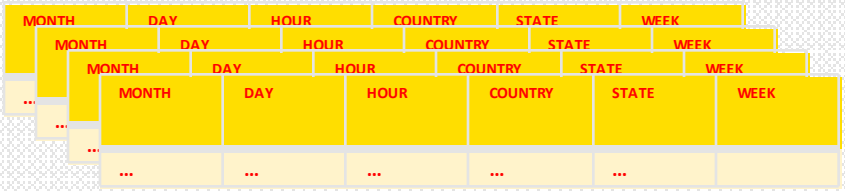
Dimension	Level	Level	Level	Level
TIME	YEAR	MONTH	DAY	HOUR
TIME_WEEKLY	YEAR	DAY_NAME		
GEO	COUNTRY	STATE		
PUBLISHER	PUBLISHER	SITE_NAME	SITE_TYPE	
PUBLISHER_PLACEMENT	SITE_NAME	PLACEMENT		
CAMPAIGN	CAMPAIGN_NAME	CAMPAIGN_DESC	INDUSTRY_SEGMENT	
AD	SIZE	HEIGHT	WIDTH	
ADVERTISER	ADVERTISER_NAME			
FLIGHT	FLIGHT_NAME	FLIGHT_START_DATE	FLIGHT_END_DATE	FLIGHT_CREATIVE_ID
RID				
USERID	GENDER	AGE_BUCKET	AGE	

jovianDATA

MarketShare

Materializing the Cube

YEAR	MONTH	DAY	HOUR	COUNTRY	STATE	WEEK	PUBLISHER	SITE_NAME	CAMPAIGN_NAME	CAMPAIGN_DESC	SIZE	HEIGHT	FLIGHT_NAME	FLIGHT_START_DATE	FLIGHT_END_DATE	FLIGHT_CREATIVE_ID
2009	June	3	1	US	CA	21	AOL	AOL Autos	Ford	Spring	2x2	10		11/06	11/08	Banner



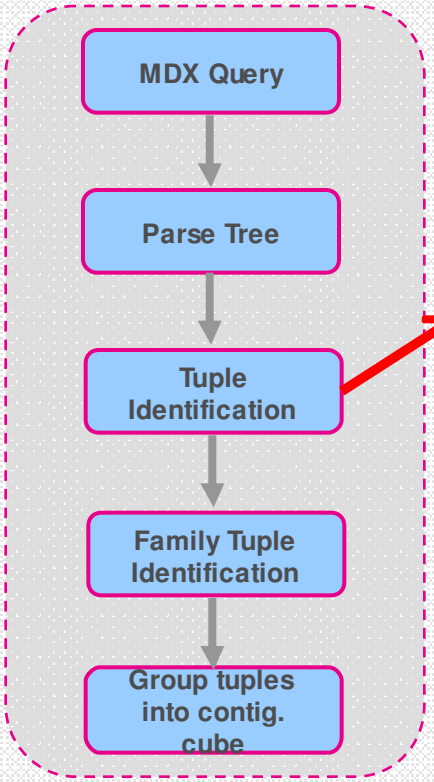
Generate 2^6 combinations for each fast changing dimension
Major data explosion here – so use more nodes

Hash Partition on Fast Dim Vector (Maintain balance via SLA)

Consolidate the fast partition vectors for all tuples
Heavy duplicate elimination here

MarketShare

Life of an MDX Query




```
Union
 (
  Union
  (
   Union
   (
    CrossJoin
    (
     {[Promotion Media].[All Media]}
     ,{[Product].[All Products]}
    )
    ,CrossJoin
    (
     {[Promotion Media].[All Media]}
     ,[Product].[All Products].Children
    )
   )
   ,CrossJoin
   (
    {[Promotion Media].[All Media]}
    ,[Product].[All Products].[Drink].Children
   )
  )
  ,Union
  (
   Union
   (
    CrossJoin
    (
     [Promotion Media].[All Media].Children
     ,{[Product].[All Products]}
    )
    ,CrossJoin
    (
     [Promotion Media].[All Media].Children
     ,[Product].[All Products].Children
    )
   )
   ,CrossJoin
   (
    [Promotion Media].[All Media].Children
    ,[Product].[All Products].[Drink].Children
   )
  )
 ) ON ROWS
FROM [Sales]
WHERE [Time].[1997];
```

Tuple Set Generation from complex trees of CrossJoin, Union, Predicates etc.

Convert incoming query to a set of multi-dimensional tuples

MarketShare

Tuple Access Layer

YEAR	MONTH	DAY	HOUR	COUNTRY	WEEK	STATE	PUBLISHER	SITE_NAME	CAMPAIGN_NAME	CAMPAIGN_DESC	SIZE	HEIGHT	FLIGHT_NAME	FLIGHT_START_DATE	FLIGHT_END_DATE	FLIGHT_CREATIVE_ID
2009	June	3	1	US	*	*	AOL Autos	*	*	*	*	*	Ford	*	*	Banner

MONTH	DAY	HOUR	COUNTRY	STATE	WEEK
June	3	1	US	*	*

Locate the partition for the fast dimension values

Note here that STATE is set to '*' but it has already been materialized
Therefore, we eliminate any sum across all states

YEAR	PUBLISHER	SITE_NAME	CAMPAIGN_NAME	CAMPAIGN_DESC	SIZE	HEIGHT	FLIGHT_NAME	FLIGHT_START_DATE	FLIGHT_END_DATE	FLIGHT_CREATIVE_ID
2009	AOL Autos	*	*	*	*	*	Ford	*	*	Banner

On the local partition, do the aggregation to calculate the '*'

MarketShare

5 types of partitions maintained in the cloud

Exclusive

EC2 nodes that are allocated for specific keys

Permanent

EC2 nodes that are permanently allocated to service queries

Temporary

EC2 nodes that host temporary replicas

Archive

S3 storage that is not accessible directly by the query engine

Intermittent

EC2 nodes that are allocated by the loading engine

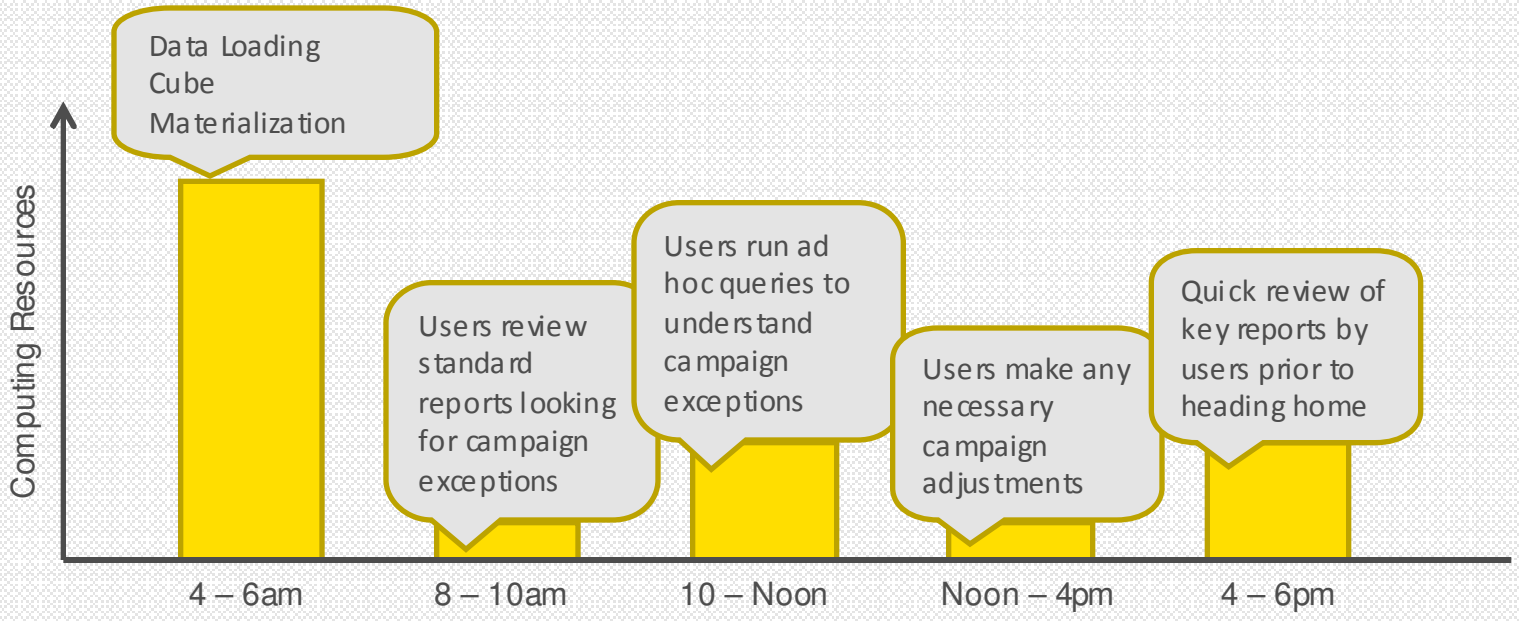
MarketShare

Taxi Meter
Reducing permanent resources

MarketShare

Usage Patterns

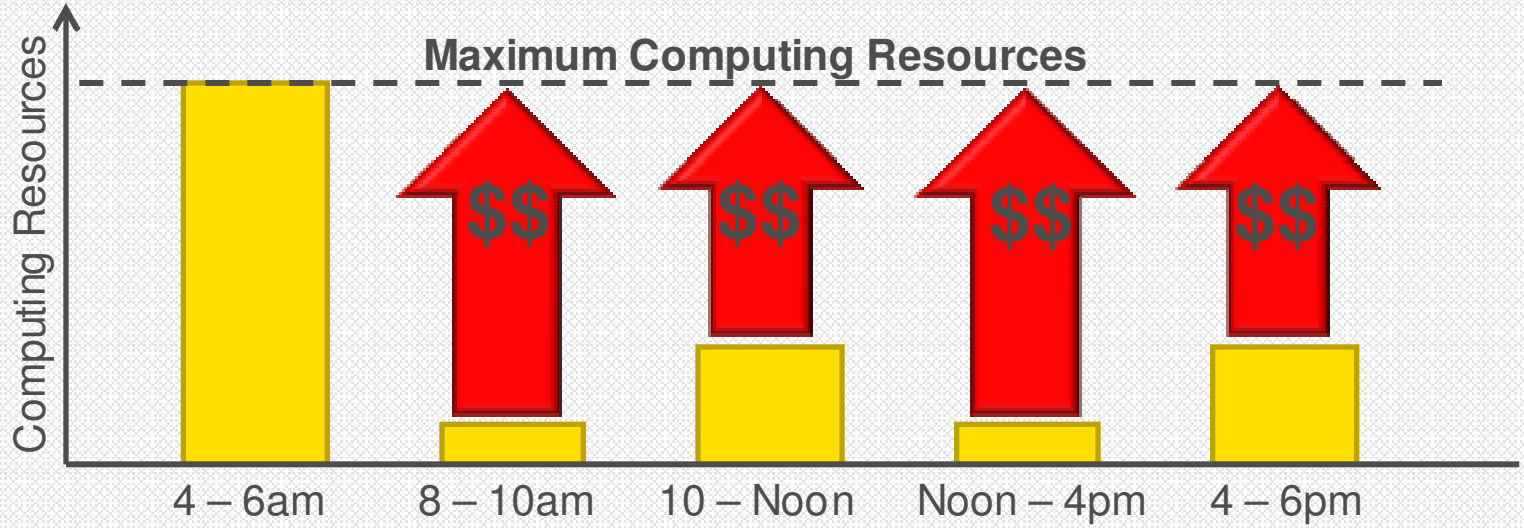
- Usage patterns vary throughout the day and throughout the week
- A couple of periods of heavy usage daily, followed by moderate to low usage



MarketShare

Traditional Computing Approach

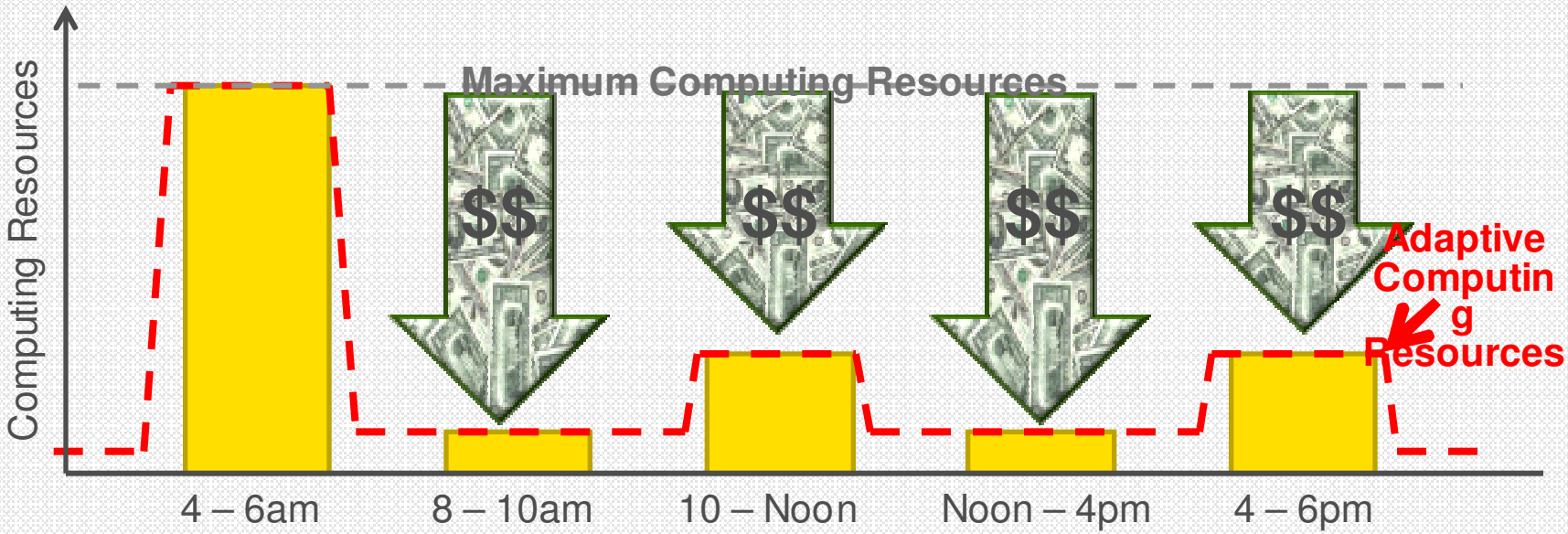
- Traditional computing approach buys enough computing resources to meet peak usage demand
- Even many cloud "solutions" provide only the peak computing power option with no way to dynamically reallocate the computing resources to match the current usage demand
- **Result:** Substantial waste in computing resources and money



MarketShare

“Adaptive” Computing Economics

- Finely matching computing resources to user usage patterns can provide a 50% to 90% cost savings versus the traditional computing resource allocation approach
  - **Result**: Lower cost **with** improvements in availability and performance

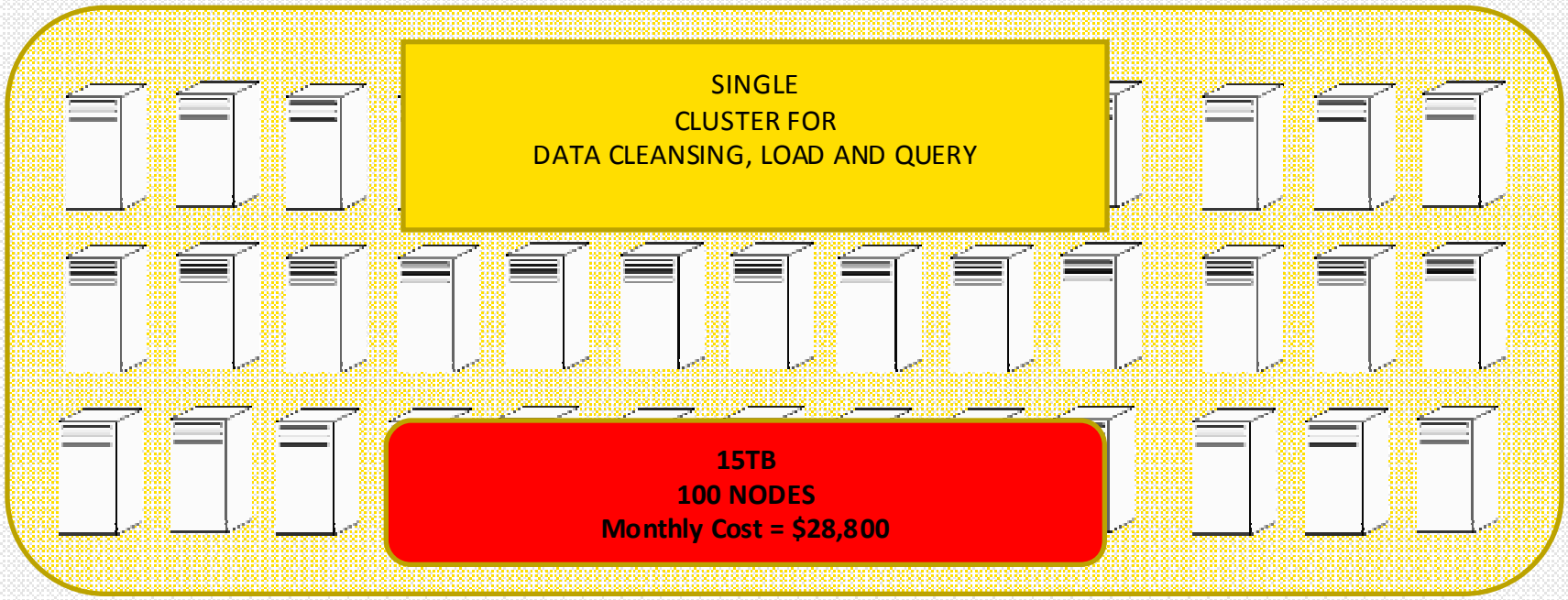


MarketShare

Intermittent scalability
Using large number of nodes during load time

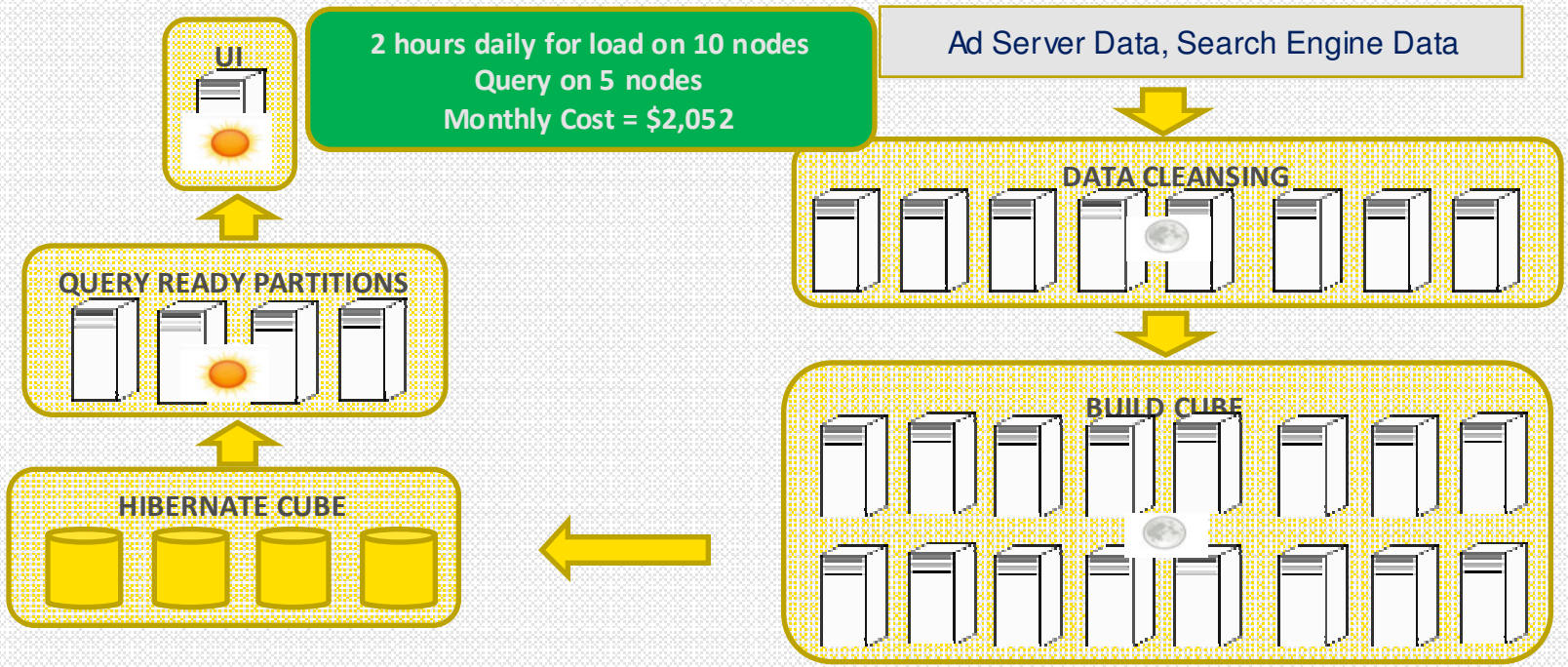
MarketShare

Managing CapEx with Role Based Clusters



MarketShare

Role Based Clusters



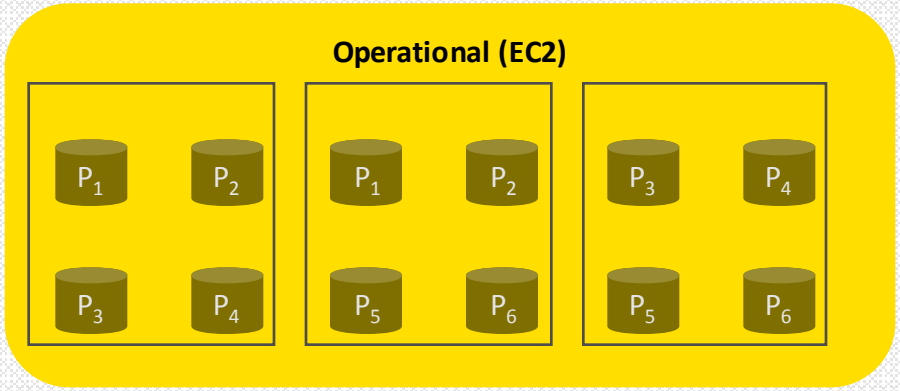
MarketShare

Selective replication for hot partitions

MarketShare

Partition level query slowdown

- Dynamic statistics
  - The query execution system logs status for each partition
  - If a particular partition is regularly lagging behind, it is marked for replication
- Static statistics
  - The query execution system identifies skews in specific partitions
  - Partitions with size skew etc are marked for replication

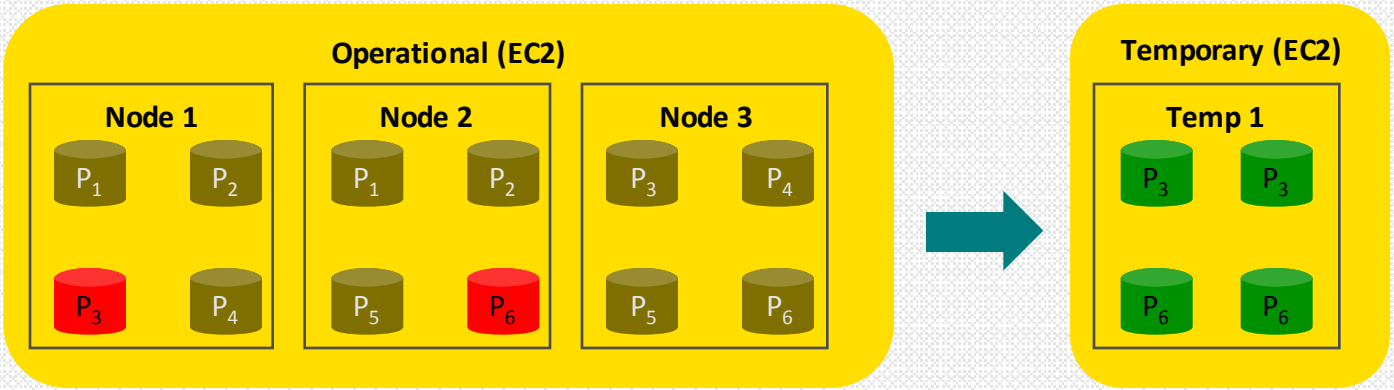


Partition	Size	Average Execution Time
P_1	1MB	1.2s
P_2	2MB	
P_3	**1.5MB**	**60s**
…		

MarketShare

Fixing partition level slowdown

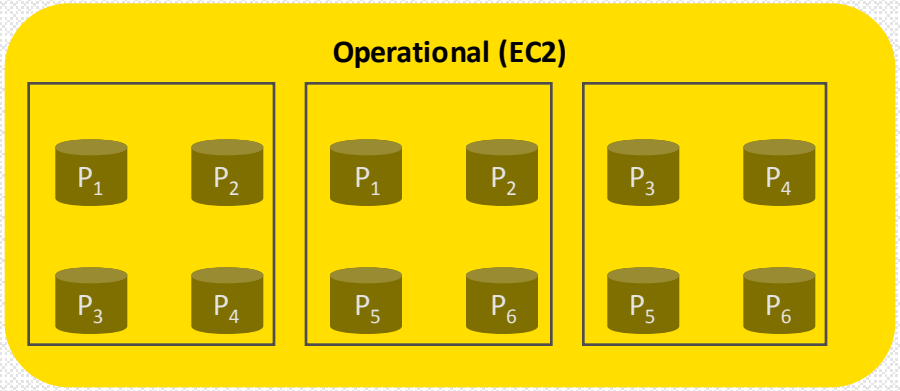
- If the query execution system detects SLA violations
- Adds two new temporary nodes (Temp 1)
- Creates new replicas for the ‘hot’ partitions



MarketShare

Key level query slowdown

- Key Level Dynamic statistics
  - A particular key takes time for materializing various facets of the cube

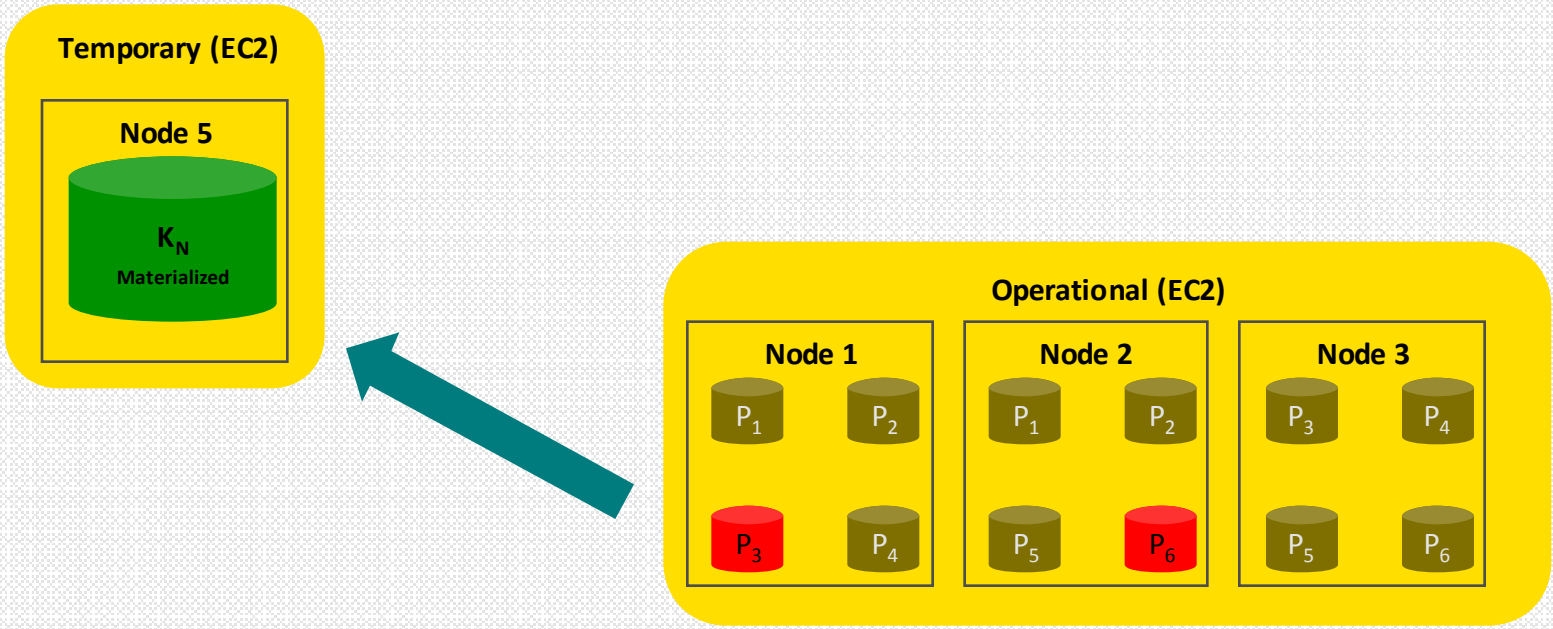


Keys	Size	Average Execution Time
P_3. K_1	20MB	120s
…		

MarketShare

Fixing partition level slowdown

- If the query execution system detects SLA violations for a particular key
- Adds a new temporary node (Temp 2)
- Denormalizes the key such that all data for that key is materialized



MarketShare

Partitions can be in 5 different states