Carnegie Mellon University

MAKE YOUR DATABASE DREAM OF ELECTRIC SHEEP

DESIGNING FOR AUTONOMOUS OPERATION

@andy_pavlo

Part #1 – Background
Part #2 – Engineering
Part #3 – Oracle Rant

1970-1990s

Self-Adaptive Databases

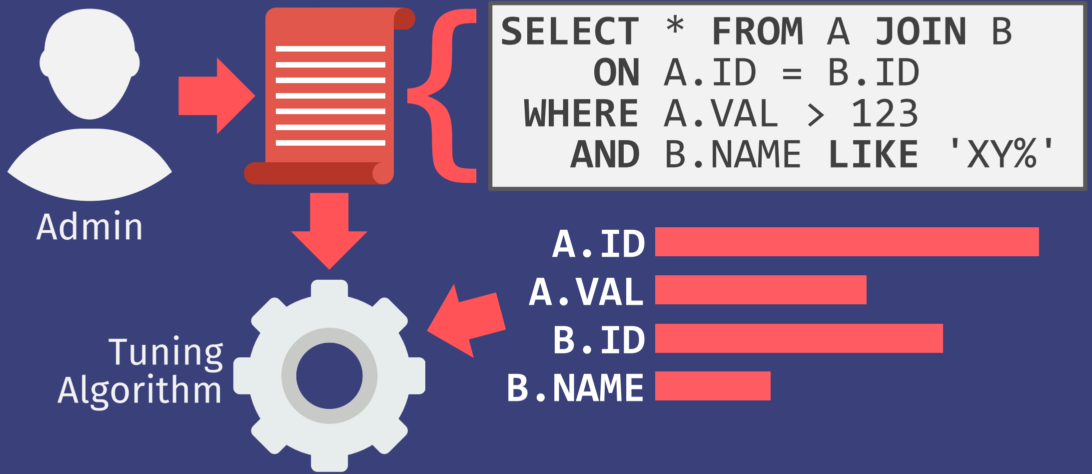
- → Index Selection
- → Partitioning / Sharding
- → Data Placement

1970-1990s
Self-Adaptive Databases

Admin

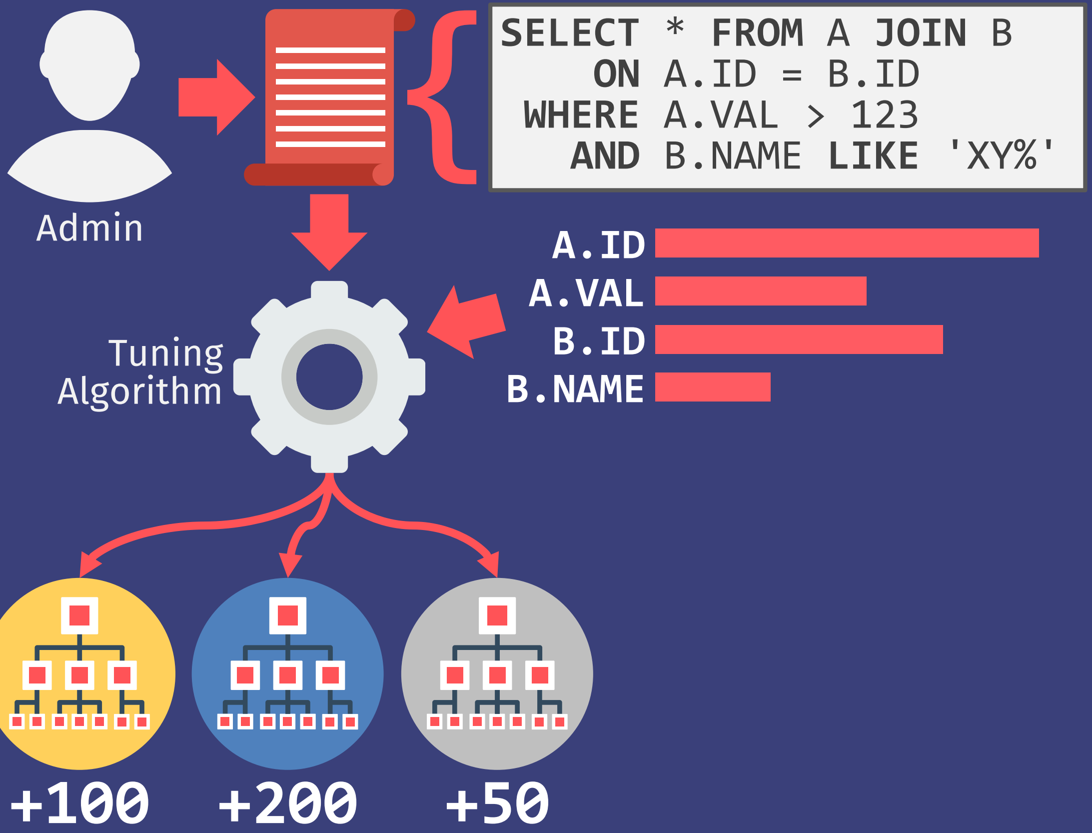
```
SELECT * FROM A JOIN B
    ON A.ID = B.ID
 WHERE A.VAL > 123
   AND B.NAME LIKE 'XY%'
```

1970-1990s Self-Adaptive Databases



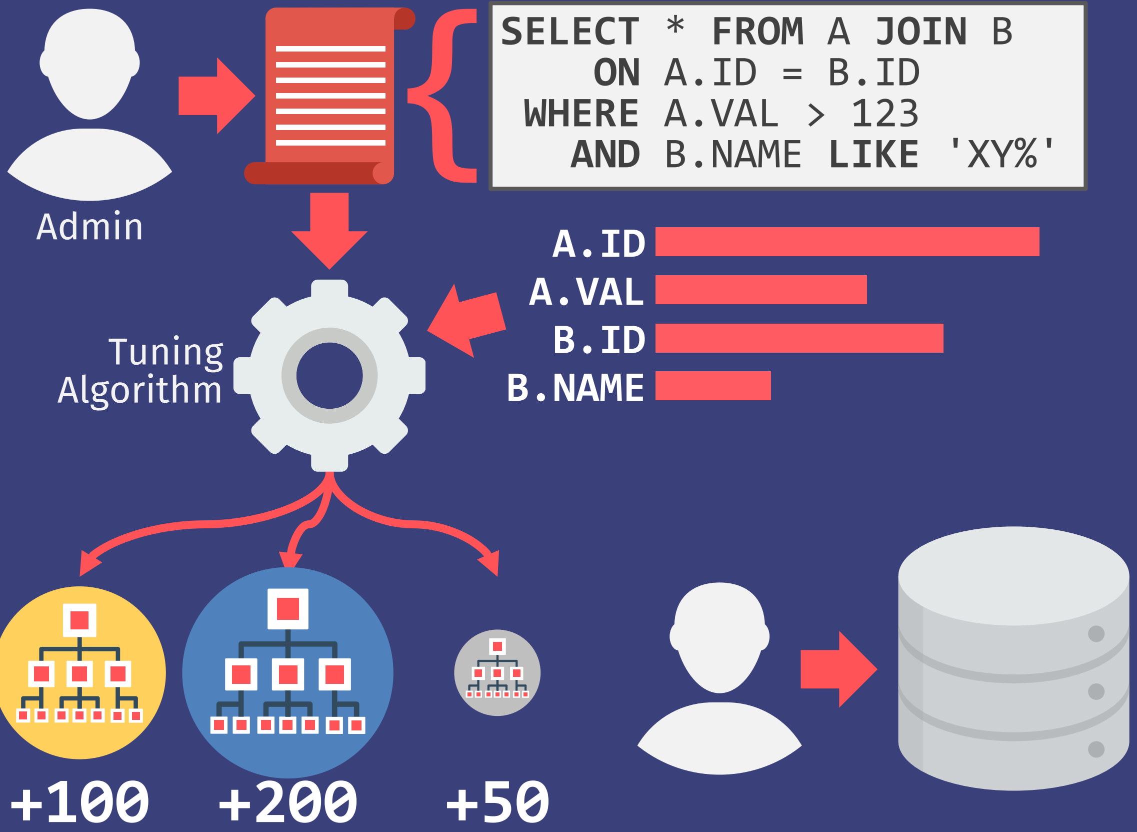
1970-1990s

Self-Adaptive Databases



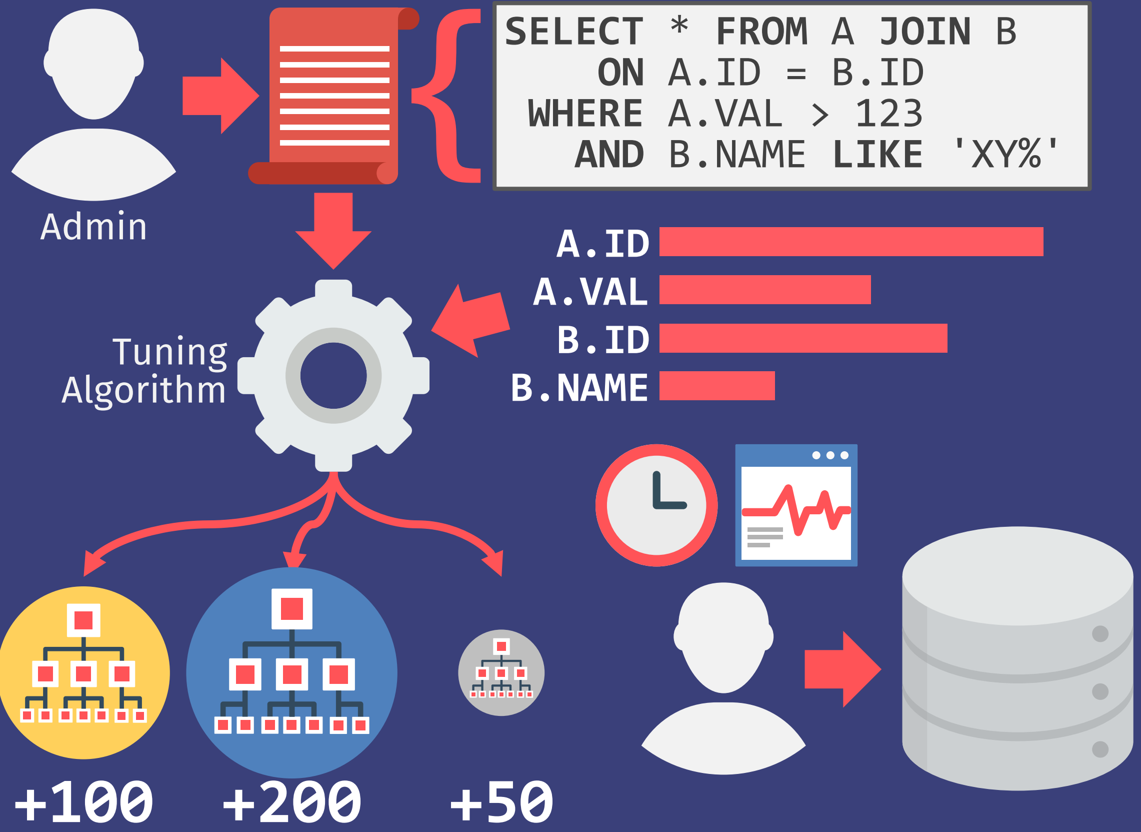
1970-1990s

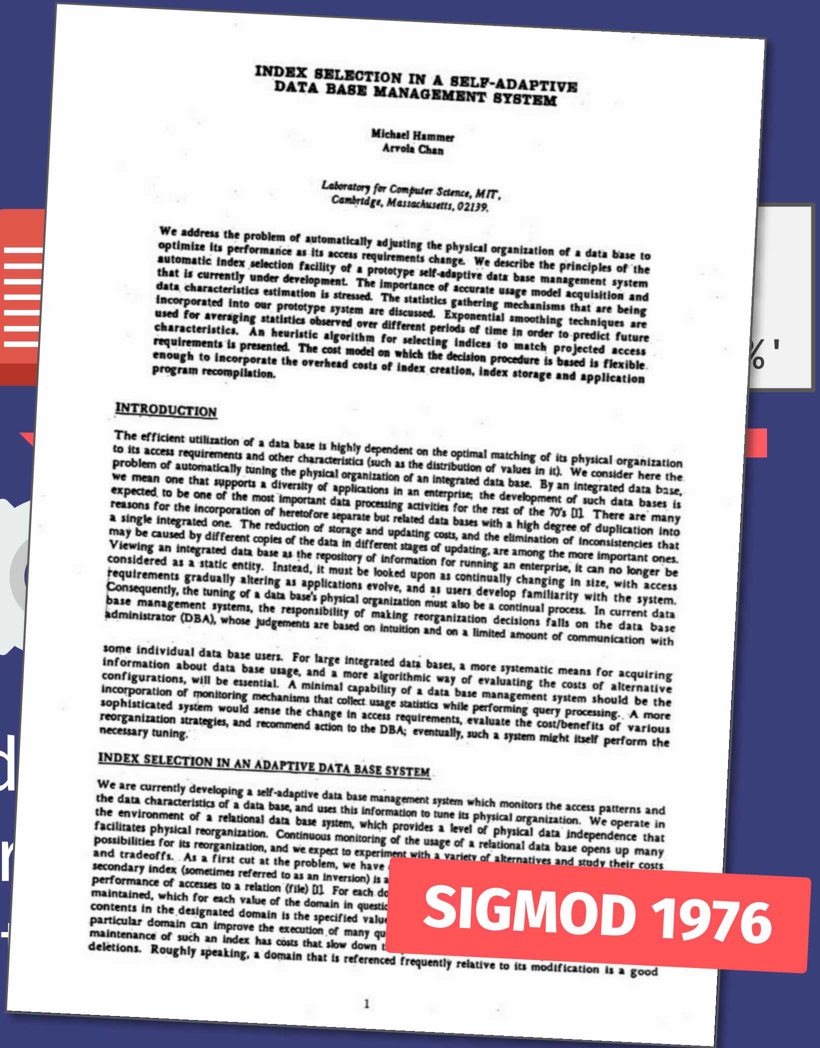
Self-Adaptive Databases



AUTONOMOUS DBMSs
SELF-ADAPTIVE DATABASES

1970-1990s
Self-Adaptive Databases





INDEX SELECTION IN A SELF-ADAPTIVE DATA BASE MANAGEMENT SYSTEM

Michael Hammer
Arvola Chan

Laboratory for Computer Science, MIT,
Cambridge, Massachusetts, 02139.

We address the problem of automatically adjusting the physical organization of a data base to optimize its performance as its access requirements change. We describe the principles of the automatic index selection facility of a prototype self-adaptive data base management system that is currently under development. The importance of accurate usage model acquisition and data characteristics estimation is stressed. The statistics gathering mechanisms that are being incorporated into our prototype system are discussed. Exponential smoothing techniques are used for averaging statistics observed over different periods of time in order to predict future characteristics. An heuristic algorithm for selecting indices to match projected access requirements is presented. The cost model on which the decision procedure is based is flexible enough to incorporate the overhead costs of index creation, index storage and application program recompilation.

INTRODUCTION

The efficient utilization of a data base is highly dependent on the optimal matching of its physical organization to its access requirements and other characteristics (such as the distribution of values in it). We consider here the problem of automatically tuning the physical organization of an integrated data base. By an integrated data base, we mean one that supports a diversity of applications in an enterprise; the development of such data bases is expected to be one of the most important data processing activities for the rest of the 70's [1]. There are many reasons for the incorporation of heretofore separate but related data bases with a high degree of duplication into a single integrated one. The reduction of storage and updating costs, and the elimination of inconsistencies that may be caused by different copies of the data in different stages of updating, are among the more important ones. Viewing an integrated data base as the repository of information for running an enterprise, it can no longer be considered as a static entity. Instead, it must be looked upon as continually changing in size, with access requirements gradually altering as applications evolve, and as users develop familiarity with the system. Consequently, the tuning of a data base's physical organization must also be a continual process. In current data base management systems, the responsibility of making reorganization decisions falls on the data base administrator (DBA), whose judgements are based on intuition and on a limited amount of communication with

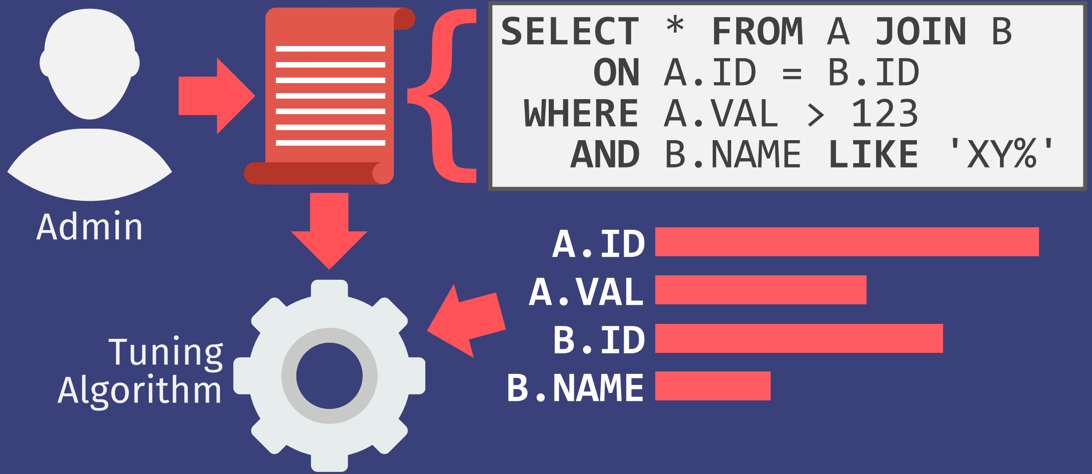
some individual data base users. For large integrated data bases, a more systematic means for acquiring information about data base usage, and a more algorithmic way of evaluating the costs of alternative configurations, will be essential. A minimal capability of a data base management system should be the incorporation of monitoring mechanisms that collect usage statistics while performing query processing. A more sophisticated system would sense the change in access requirements, evaluate the cost/benefits of various reorganization strategies, and recommend action to the DBA; eventually, such a system might itself perform the necessary tuning.

INDEX SELECTION IN AN ADAPTIVE DATA BASE SYSTEM

We are currently developing a self-adaptive data base management system which monitors the access patterns and the data characteristics of a data base, and uses this information to tune its physical organization. We operate in the environment of a relational data base system, which provides a level of physical data independence that facilitates physical reorganization. Continuous monitoring of the usage of a relational data base opens up many possibilities for its reorganization, and we expect to experiment with a variety of alternatives and study their costs and tradeoffs. As a first cut at the problem, we have secondary index (sometimes referred to as an inversion) is a performance of accesses to a relation (file) [1]. For each do maintained, which for each value of the domain in questi contents in the designated domain is the specified valu particular domain can improve the execution of many qu maintenance of such an index has costs that slow down t deletions. Roughly speaking, a domain that is referenced frequently relative to its modification is a good

1

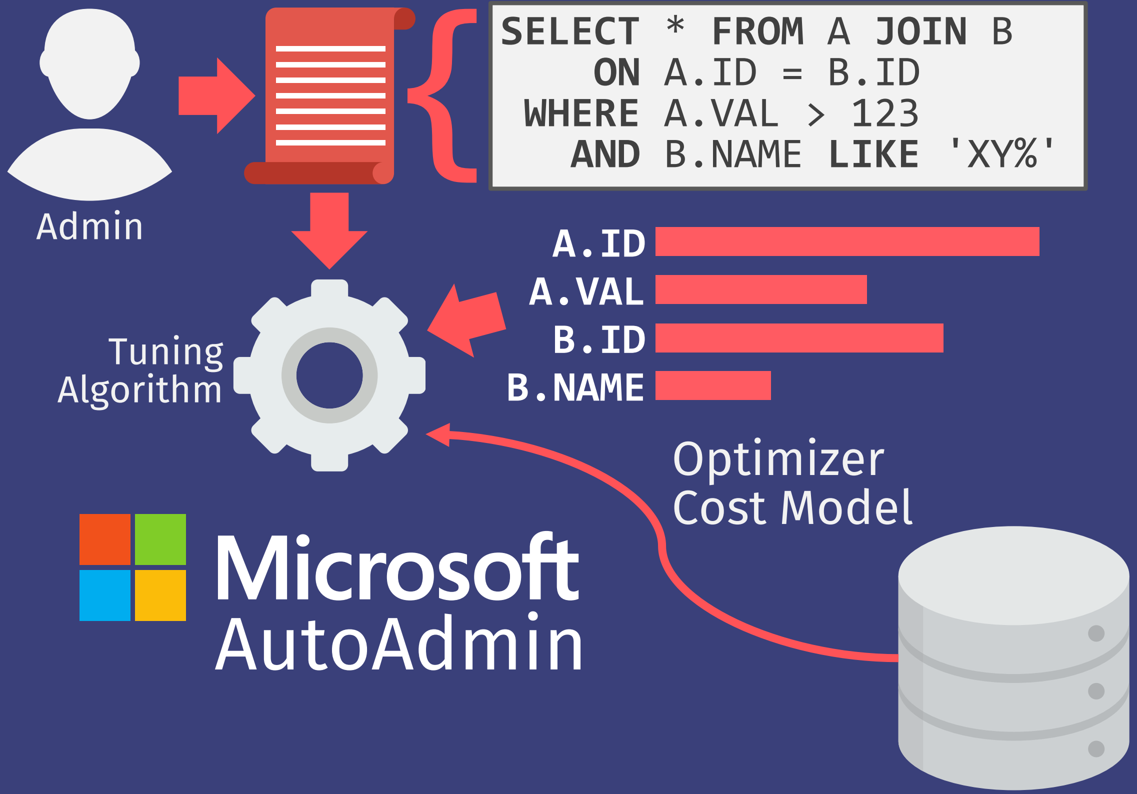
1990-2000s
Self-Tuning Databases



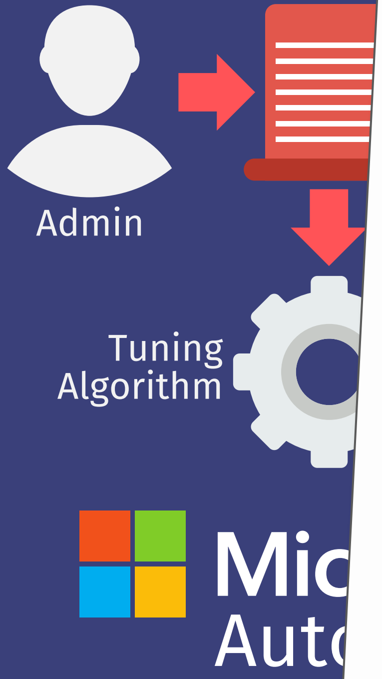
→Index Selection
→Partitioning / Sharding
→Data Placement

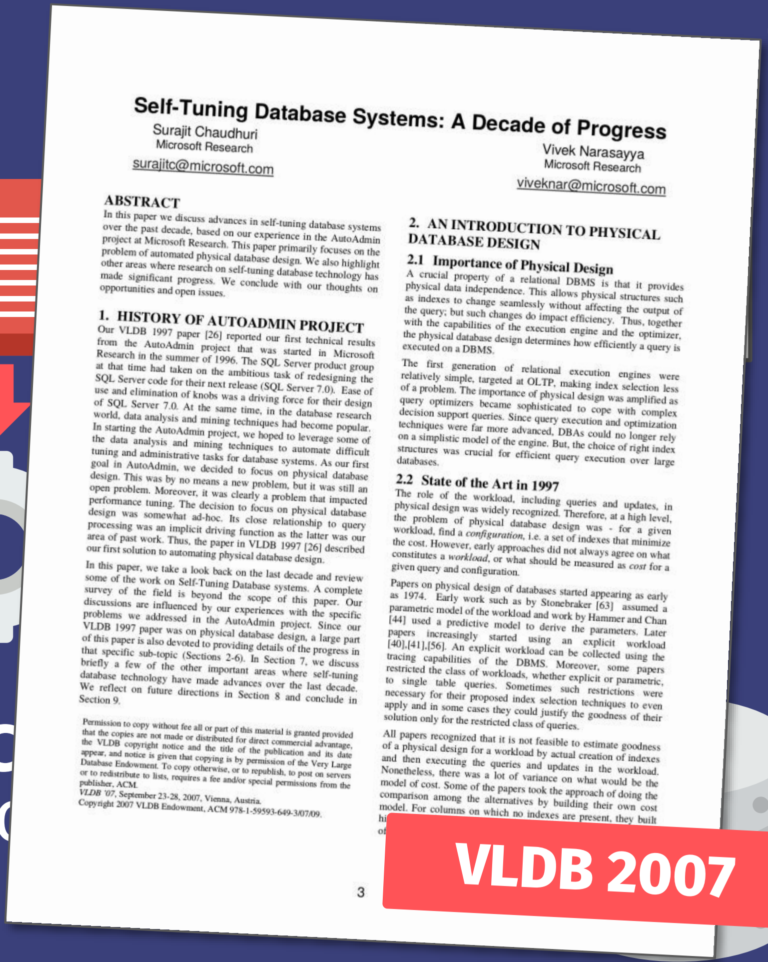
1990-2000s

Self-Tuning Databases



1990-2000s Self-Tuning Databases





Self-Tuning Database Systems: A Decade of Progress

Surajit Chaudhuri
Microsoft Research
surajitc@microsoft.com

Vivek Narasayya
Microsoft Research
viveknar@microsoft.com

ABSTRACT

In this paper we discuss advances in self-tuning database systems over the past decade, based on our experience in the AutoAdmin project at Microsoft Research. This paper primarily focuses on the problem of automated physical database design. We also highlight other areas where research on self-tuning database technology has made significant progress. We conclude with our thoughts on opportunities and open issues.

1. HISTORY OF AUTOADMIN PROJECT

Our VLDB 1997 paper [26] reported our first technical results from the AutoAdmin project that was started in Microsoft Research in the summer of 1996. The SQL Server product group at that time had taken on the ambitious task of redesigning the SQL Server code for their next release (SQL Server 7.0). Ease of use and elimination of knobs was a driving force for their design of SQL Server 7.0. At the same time, in the database research world, data analysis and mining techniques had become popular. In starting the AutoAdmin project, we hoped to leverage some of the data analysis and mining techniques to automate difficult tuning and administrative tasks for database systems. As our first goal in AutoAdmin, we decided to focus on physical database design. This was by no means a new problem, but it was still an open problem. Moreover, it was clearly a problem that impacted performance tuning. The decision to focus on physical database design was somewhat ad-hoc. Its close relationship to query processing was an implicit driving function as the latter was our area of past work. Thus, the paper in VLDB 1997 [26] described our first solution to automating physical database design.

In this paper, we take a look back on the last decade and review some of the work on Self-Tuning Database systems. A complete survey of the field is beyond the scope of this paper. Our discussions are influenced by our experiences with the specific problems we addressed in the AutoAdmin project. Since our VLDB 1997 paper was on physical database design, a large part of this paper is also devoted to providing details of the progress in that specific sub-topic (Sections 2-6). In Section 7, we discuss briefly a few of the other important areas where self-tuning database technology have made advances over the last decade. We reflect on future directions in Section 8 and conclude in Section 9.

VLDB '07, September 23-28, 2007, Vienna, Austria.

2. AN INTRODUCTION TO PHYSICAL DATABASE DESIGN

2.1 Importance of Physical Design

A crucial property of a relational DBMS is that it provides physical data independence. This allows physical structures such as indexes to change seamlessly without affecting the output of the query; but such changes do impact efficiency. Thus, together with the capabilities of the execution engine and the optimizer, the physical database design determines how efficiently a query is executed on a DBMS.

The first generation of relational execution engines were relatively simple, targeted at OLTP, making index selection less of a problem. The importance of physical design was amplified as query optimizers became sophisticated to cope with complex decision support queries. Since query execution and optimization techniques were far more advanced, DBAs could no longer rely on a simplistic model of the engine. But, the choice of right index structures was crucial for efficient query execution over large databases.

2.2 State of the Art in 1997

The role of the workload, including queries and updates, in physical design was widely recognized. Therefore, at a high level, the problem of physical database design was - for a given workload, find a *configuration*, i.e. a set of indexes that minimize the cost. However, early approaches did not always agree on what constitutes a *workload*, or what should be measured as *cost* for a given query and configuration.

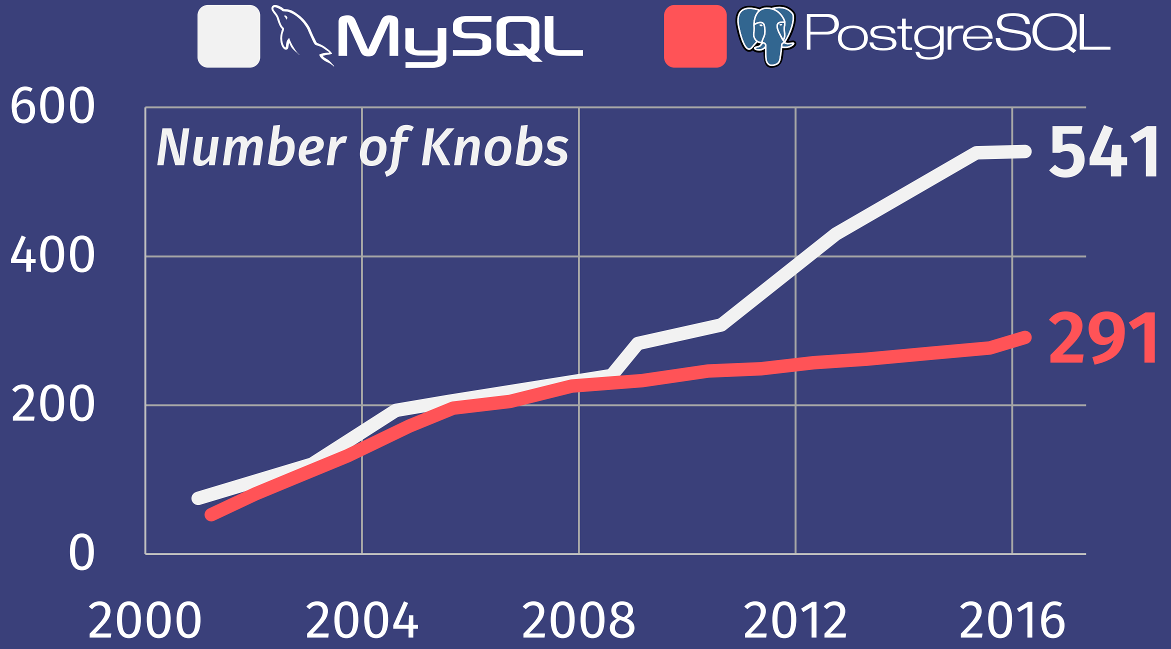
Papers on physical design of databases started appearing as early as 1974. Early work such as by Stonebraker [63] assumed a parametric model of the workload and work by Hammer and Chan [44] used a predictive model to derive the parameters. Later papers increasingly started using an explicit workload [40],[41],[56]. An explicit workload can be collected using the tracing capabilities of the DBMS. Moreover, some papers restricted the class of workloads, whether explicit or parametric, to single table queries. Sometimes such restrictions were necessary for their proposed index selection techniques to even apply and in some cases they could justify the goodness of their solution only for the restricted class of queries.

All papers recognized that it is not feasible to estimate goodness of a physical design for a workload by actual creation of indexes and then executing the queries and updates in the workload. Nonetheless, there was a lot of variance on what would be the model of cost. Some of the papers took the approach of doing the comparison among the alternatives by building their own cost model. For columns on which no indexes are present, they built

VLDB 2007

1990-2000s

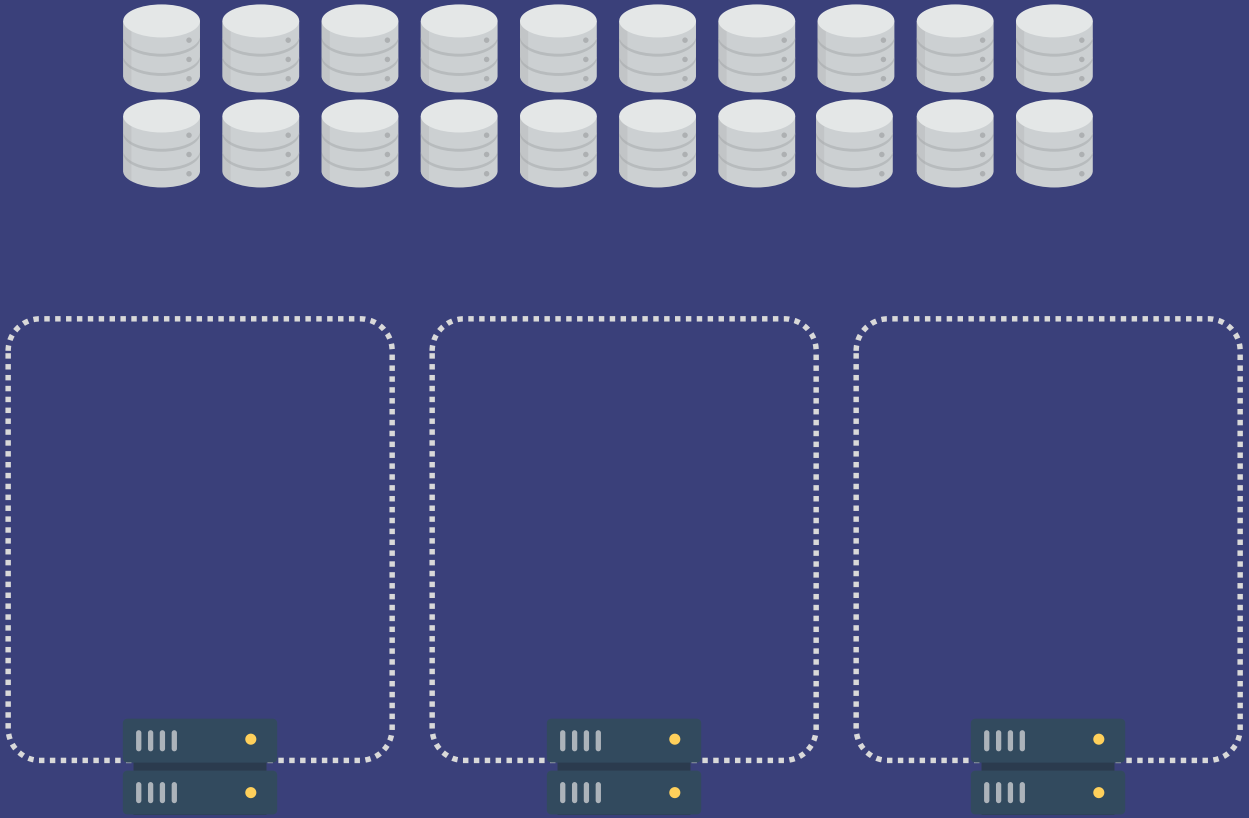
Self-Tuning Databases



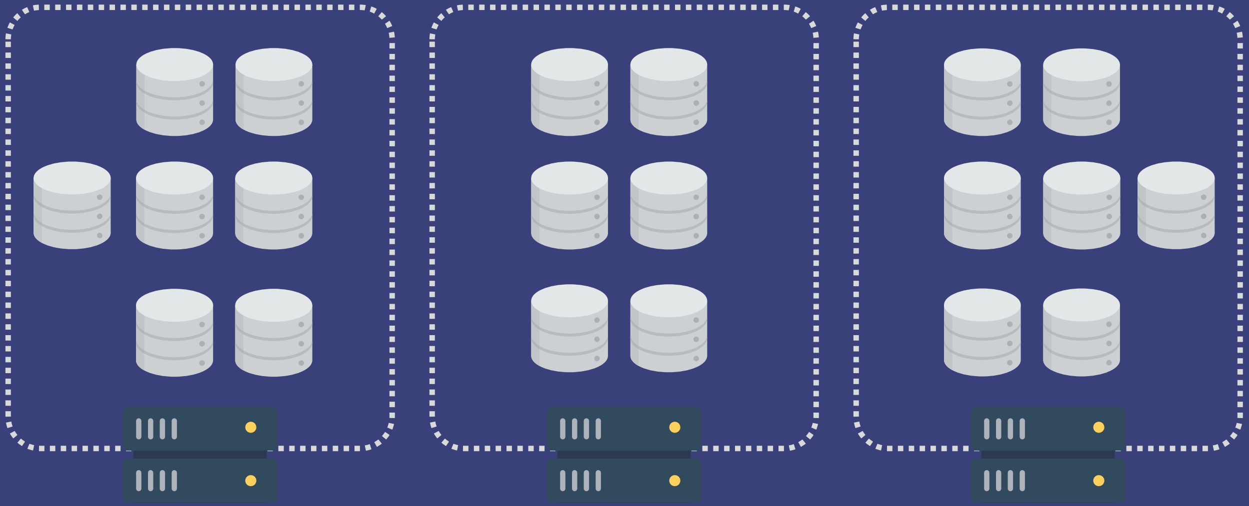
→Knob Configuration

2010s

Cloud Databases

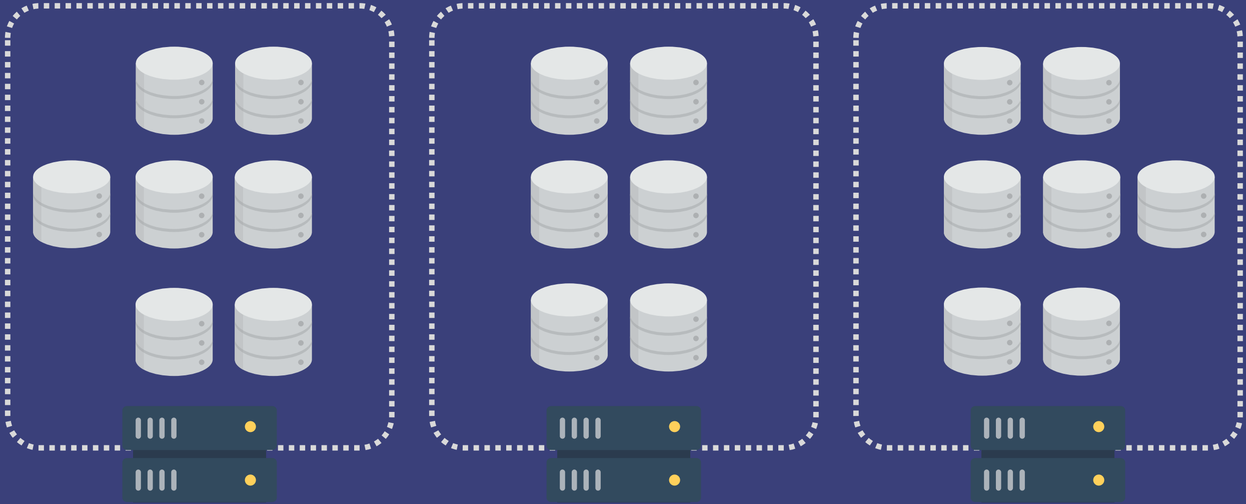


2010s
Cloud Databases



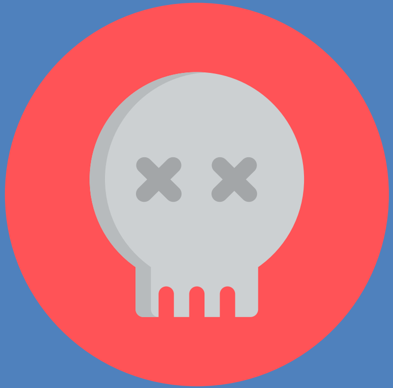
2010s
Cloud Databases

- → Initial Placement
- → Tenant Migration

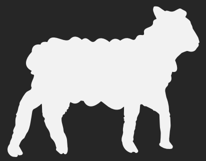


Why is this previous work insufficient?

Problem #1
Human Judgements



Problem #2
Reactionary Measures



What is different this time?

Better hardware.
Better machine learning tools.
Better appreciation for data.

We seek to complete the circle in autonomous databases.

OtterTune

Existing Systems

Peloton

New System

Database Tuning-as-a-Service

→ Automatically generate DBMS knob configurations.

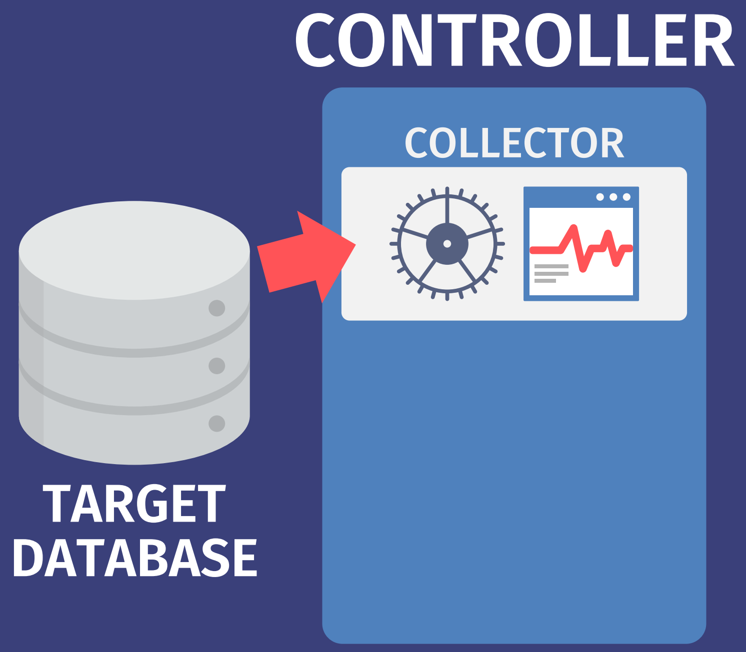
→ Reuse data from previous tuning sessions.

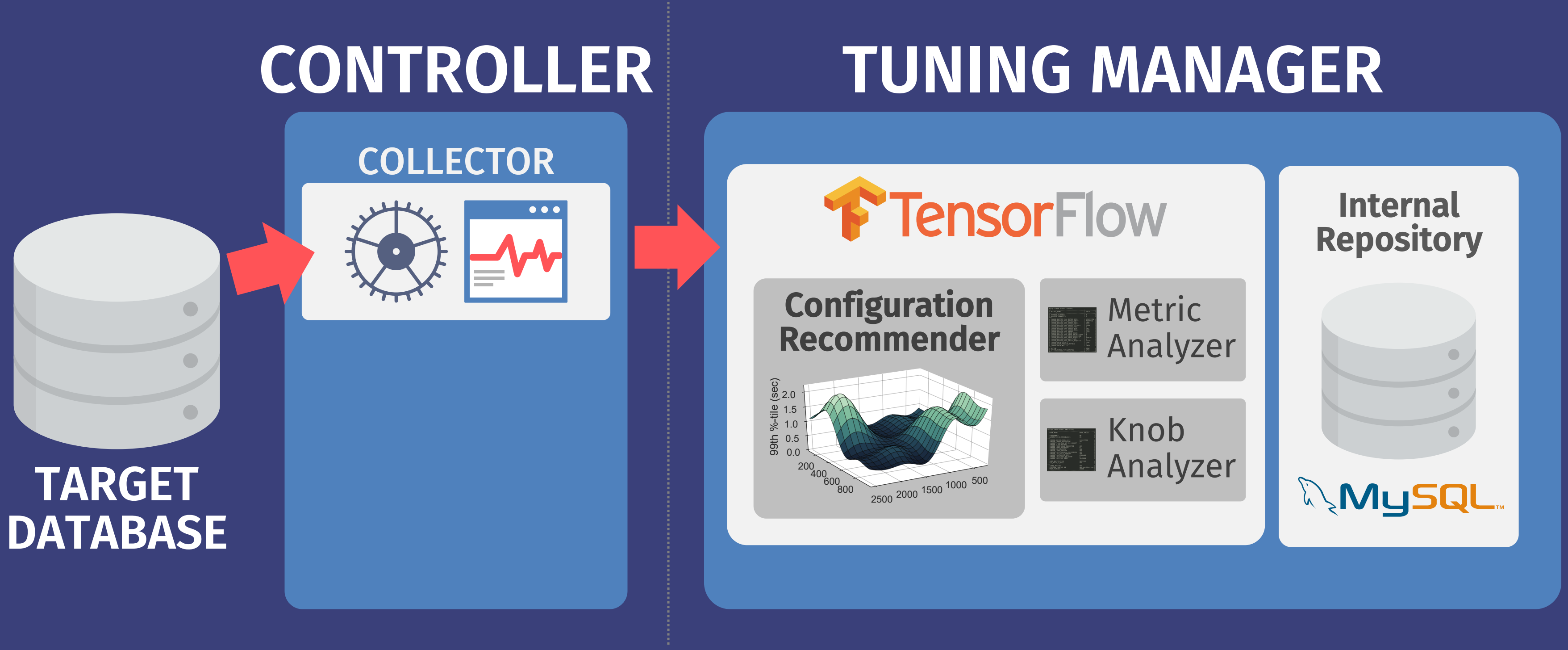
OtterTune

ottertune.cs.cmu.edu

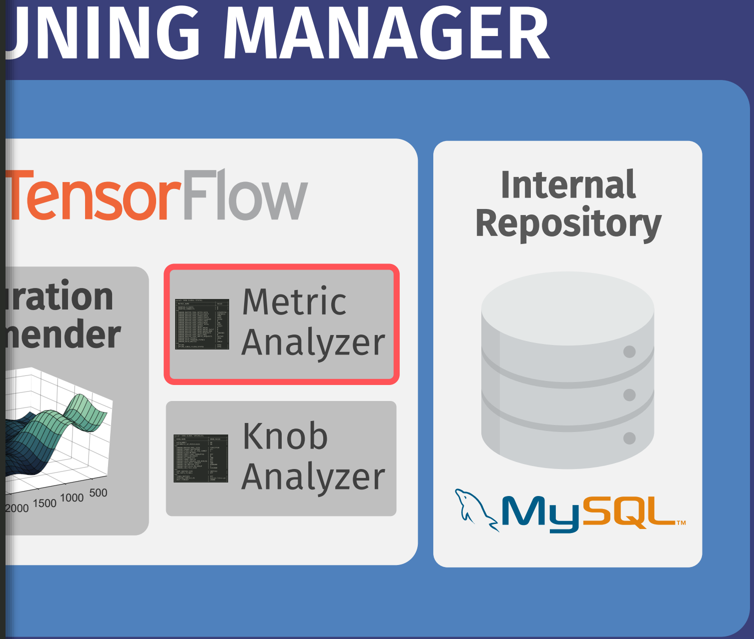
Supported Systems

PostgreSQL vectorwise MySQL Greenplum


CONTROLLER
COLLECTOR
TARGET
DATABASE

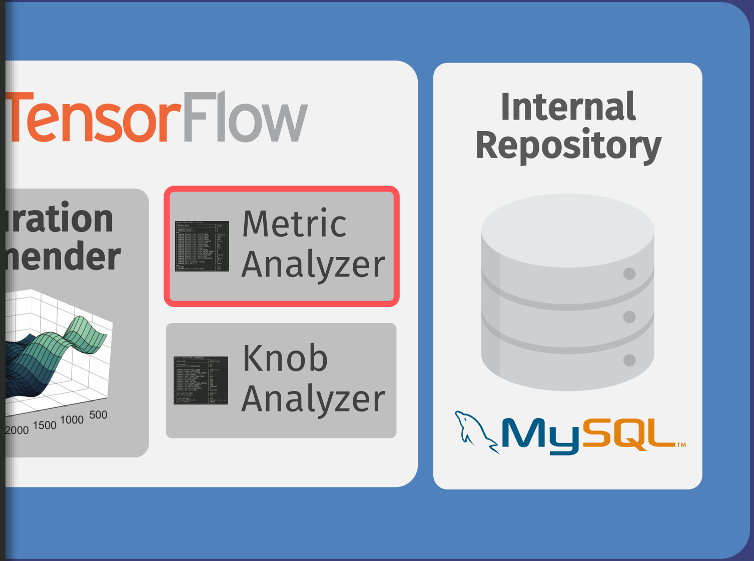

CONTROLLER
TUNING MANAGER
COLLECTOR
TensorFlow
Internal Repository
Configuration Recommender
99th %-tile (sec)
2.0
1.5
1.0
0.5
0.0
200
400
600
800
2500
2000
1500
1000
500
Metric Analyzer
Knob Analyzer
MySQL
TARGET DATABASE

```
mysql> SHOW GLOBAL STATUS;
+-----------------------------------------+------------+
| METRIC_NAME                             | VALUE      |
+-----------------------------------------+------------+
| ABORTED_CLIENTS                         | 0          |
| ABORTED_CONNECTS                        | 0          |
•••
| INNODB_BUFFER_POOL_BYTES_DATA           | 129499136  |
| INNODB_BUFFER_POOL_BYTES_DIRTY          | 76070912   |
| INNODB_BUFFER_POOL_PAGES_DATA           | 7904       |
| INNODB_BUFFER_POOL_PAGES_DIRTY          | 4643       |
| INNODB_BUFFER_POOL_PAGES_FLUSHED        | 25246      |
| INNODB_BUFFER_POOL_PAGES_FREE           | 0          |
| INNODB_BUFFER_POOL_PAGES_MISC           | 288        |
| INNODB_BUFFER_POOL_PAGES_TOTAL          | 8192       |
| INNODB_BUFFER_POOL_READS                | 15327      |
| INNODB_BUFFER_POOL_READ_AHEAD           | 0          |
| INNODB_BUFFER_POOL_READ_AHEAD_EVICT     | 0          |
| INNODB_BUFFER_POOL_READ_AHEAD_RND       | 0          |
| INNODB_BUFFER_POOL_READ_REQUESTS        | 2604302    |
| INNODB_BUFFER_POOL_WAIT_FREE            | 0          |
| INNODB_BUFFER_POOL_WRITE_REQUESTS       | 562763     |
| INNODB_DATA_FSYNCS                      | 2836       |
| INNODB_DATA_PENDING_FSYNCS              | 1          |
| INNODB_DATA_WRITES                      | 28026      |
•••
| UPTIME                                  | 5996       |
| UPTIME_SINCE_FLUSH_STATUS               | 5996       |
+-----------------------------------------+------------+
```



```
mysql> SHOW GLOBAL STATUS;
+-------------------------------------+------------+
| METRIC_NAME                         | VALUE      |
+-------------------------------------+------------+
| ABORTED_CLIENTS                     | 0          |
| ABORTED_CONNECTS                    | 0          |
•••
| INNODB_BUFFER_POOL_BYTES_DATA       | 129499136  |
| INNODB_BUFFER_POOL_BYTES_DIRTY      | 76070912   |
| INNODB_BUFFER_POOL_PAGES_DATA       | 7904       |
| INNODB_BUFFER_POOL_PAGES_DIRTY      | 4643       |
| INNODB_BUFFER_POOL_PAGES_FLUSHED    | 25246      |
| INNODB_BUFFER_POOL_PAGES_FREE       | 0          |
| INNODB_BUFFER_POOL_PAGES_MISC       | 288        |
| INNODB_BUFFER_POOL_PAGES_TOTAL      | 8192       |
| INNODB_BUFFER_POOL_READS            | 15327      |
| INNODB_BUFFER_POOL_READ_AHEAD       | 0          |
| INNODB_BUFFER_POOL_READ_AHEAD_EVICT | 0          |
| INNODB_BUFFER_POOL_READ_AHEAD_RND   | 0          |
| INNODB_BUFFER_POOL_READ_REQUESTS    | 2604302    |
| INNODB_BUFFER_POOL_WAIT_FREE        | 0          |
| INNODB_BUFFER_POOL_WRITE_REQUESTS   | 562763     |
| INNODB_DATA_FSYNCS                  | 2836       |
| INNODB_DATA_PENDING_FSYNCS          | 1          |
| INNODB_DATA_WRITES                  | 28026      |
•••
| UPTIME                              | 5996       |
| UPTIME_SINCE_FLUSH_STATUS           | 5996       |
+-------------------------------------+------------+
```

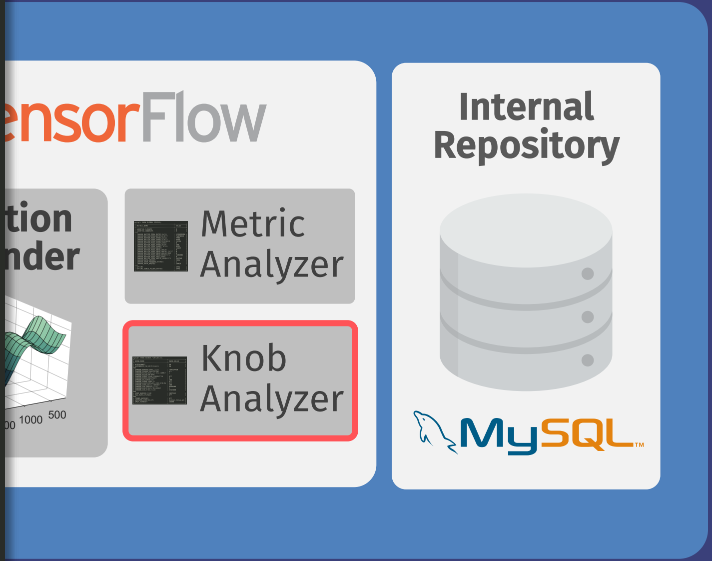
UNING MANAGER



OTTERTUNE

```
mysql> SHOW GLOBAL VARIABLES;
+-----------------------------------+------------------+
| KNOB_NAME                         | KNOB_VALUE       |
+-----------------------------------+------------------+
| AUTOCOMMIT                        | ON               |
| AUTOMATIC_SP_PRIVILEGES           | ON               |
•••
| INNODB_BUFFER_POOL_SIZE           | 134217728        |
| INNODB_CHANGE_BUFFERING           | all              |
| INNODB_FLUSH_LOG_AT_TRX_COMMIT    | 1                |
| INNODB_FLUSH_METHOD               |                  |
| INNODB_FORCE_LOAD_CORRUPTED       | OFF              |
| INNODB_FORCE_RECOVERY             | 0                |
| INNODB_IO_CAPACITY                | 200              |
| INNODB_LARGE_PREFIX               | OFF              |
| INNODB_LOCKS_UNSAFE_FOR_BINLOG    | OFF              |
| INNODB_LOCK_WAIT_TIMEOUT          | 500              |
| INNODB_LOG_BUFFER_SIZE            | 8388608          |
| INNODB_LOG_FILES_IN_GROUP         | 2                |
| INNODB_LOG_FILE_SIZE              | 5242880          |
•••
| SORT_BUFFER_SIZE                  | 2097152          |
| SQL_AUTO_IS_NULL                  | OFF              |
•••
| TIMED_MUTEXES                     | OFF              |
| VERSION_COMPILE_OS                | debian-linux-gn  |
| WAIT_TIMEOUT                      | 28800            |
+-----------------------------------+------------------+
```

NING MANAGER



```
mysql> SHOW GLOBAL VARIABLES;
+---------------------------------+-----------------+
| KNOB_NAME                       | KNOB_VALUE      |
+---------------------------------+-----------------+
| AUTOCOMMIT                      | ON              |
| AUTOMATIC_SP_PRIVILEGES         | ON              |
•••
| INNODB_BUFFER_POOL_SIZE         | 134217728       |
| INNODB_CHANGE_BUFFERING         | all             |
| INNODB_FLUSH_LOG_AT_TRX_COMMIT  | 1               |
| INNODB_FLUSH_METHOD             |                 |
| INNODB_FORCE_LOAD_CORRUPTED     | OFF             |
| INNODB_FORCE_RECOVERY           | 0               |
| INNODB_IO_CAPACITY              | 200             |
| INNODB_LARGE_PREFIX             | OFF             |
| INNODB_LOCKS_UNSAFE_FOR_BINLOG  | OFF             |
| INNODB_LOCK_WAIT_TIMEOUT        | 500             |
| INNODB_LOG_BUFFER_SIZE          | 8388608         |
| INNODB_LOG_FILES_IN_GROUP       | 2               |
| INNODB_LOG_FILE_SIZE            | 5242880         |
•••
| SORT_BUFFER_SIZE                | 2097152         |
| SQL_AUTO_IS_NULL                | OFF             |
•••
| TIMED_MUTEXES                   | OFF             |
| VERSION_COMPILE_OS              | debian-linux-gn |
| WAIT_TIMEOUT                    | 28800           |
+---------------------------------+-----------------+
```

NING MANAGER

ensorFlow

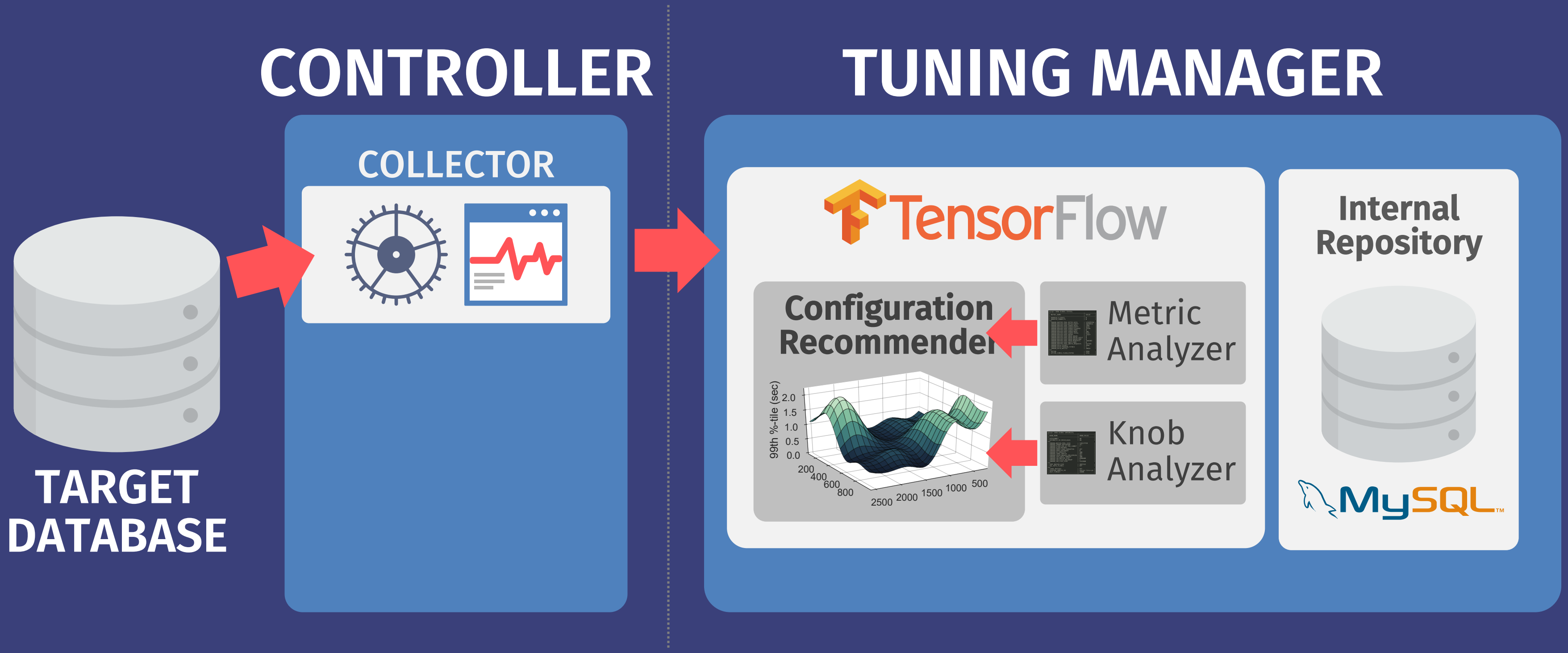
tion
nder

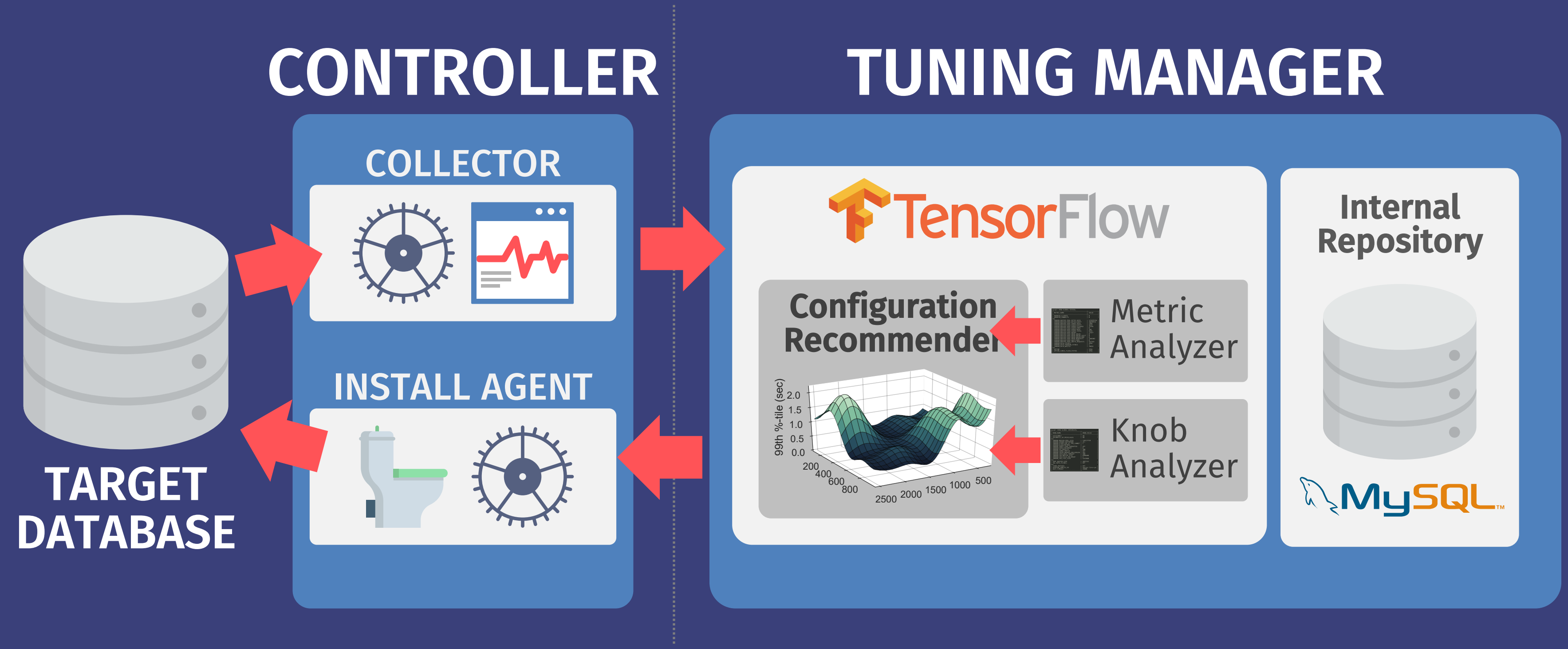
Metric Analyzer

Knob Analyzer

Internal Repository

MySQL


CONTROLLER
TUNING MANAGER
COLLECTOR
TARGET DATABASE
TensorFlow
Configuration Recommender
Metric Analyzer
Knob Analyzer
Internal Repository
MySQL
99th %-tile (sec)
2.0
1.5
1.0
0.5
0.0
200
400
600
800
2500
2000
1500
1000
500

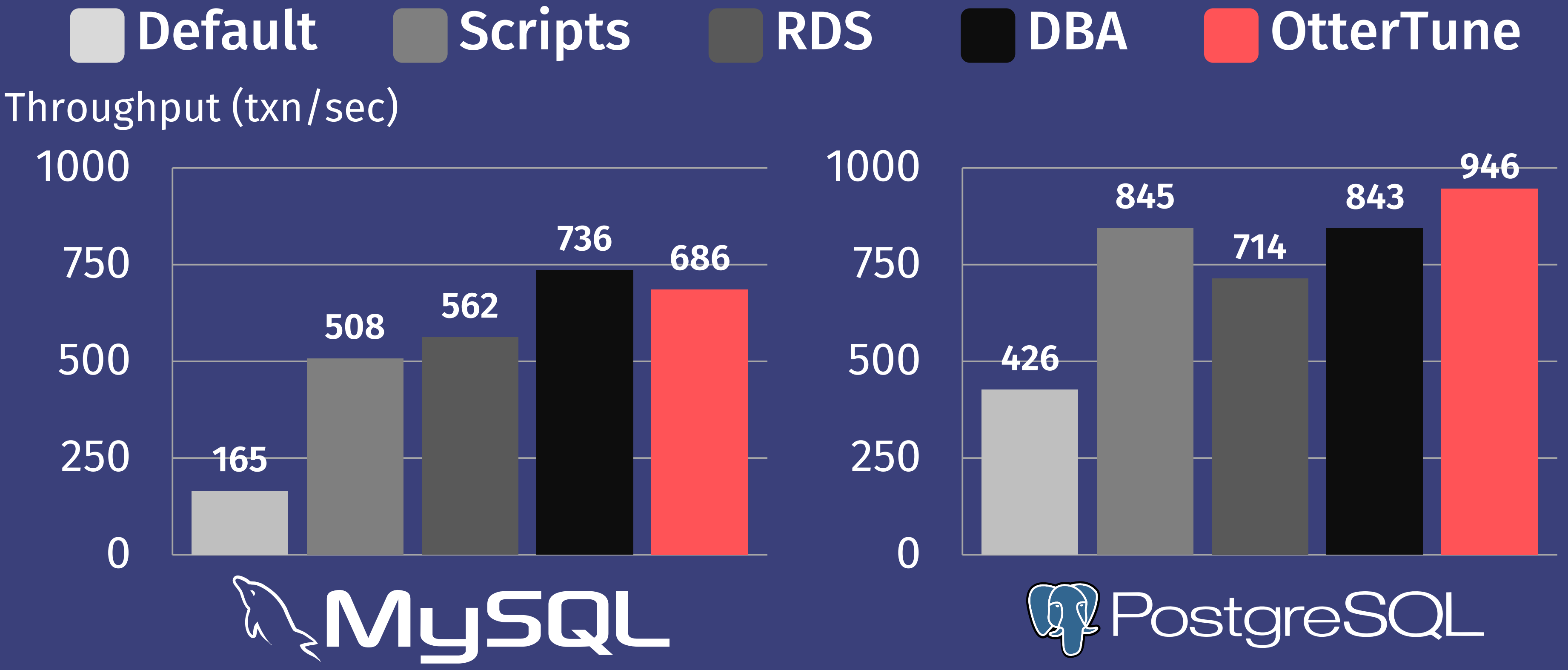

CONTROLLER
TUNING MANAGER
COLLECTOR
INSTALL AGENT
TARGET DATABASE
TensorFlow
Configuration Recommender
Metric Analyzer
Knob Analyzer
Internal Repository
MySQL

Demonstration

Postgres v9.3

TPC-C Benchmark

OTTERTUNE
TPC-C TUNING



AUTOMATIC DATABASE MANAGEMENT SYSTEM TUNING THROUGH LARGE-SCALE MACHINE LEARNING
SIGMOD 2017

Peloton

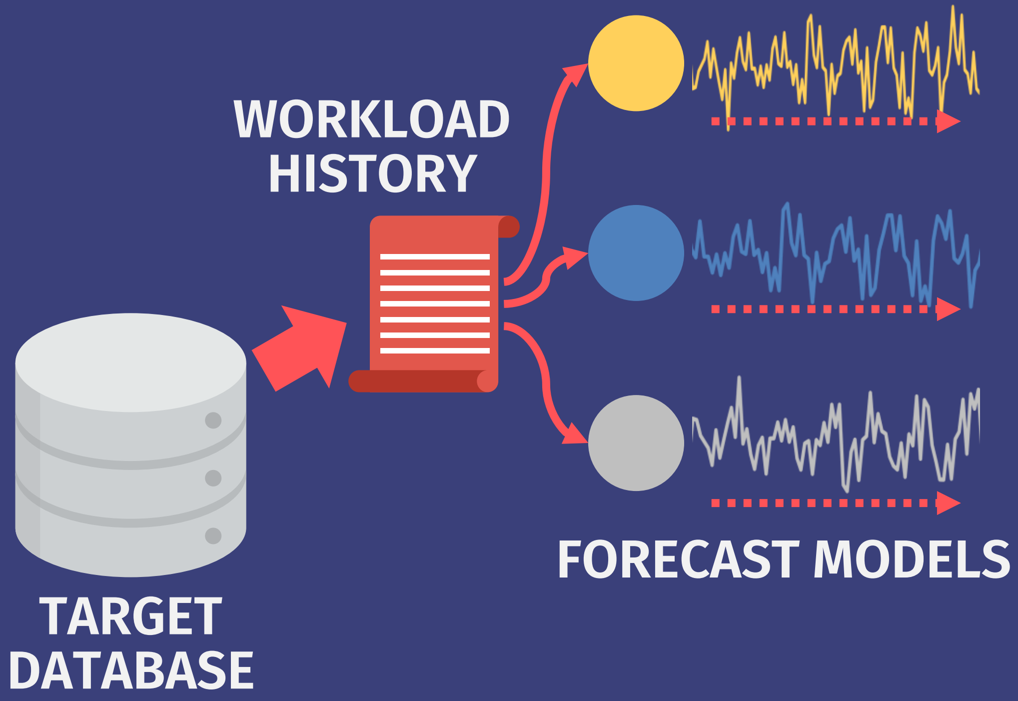
pelotondb.io

Self-Driving Database System

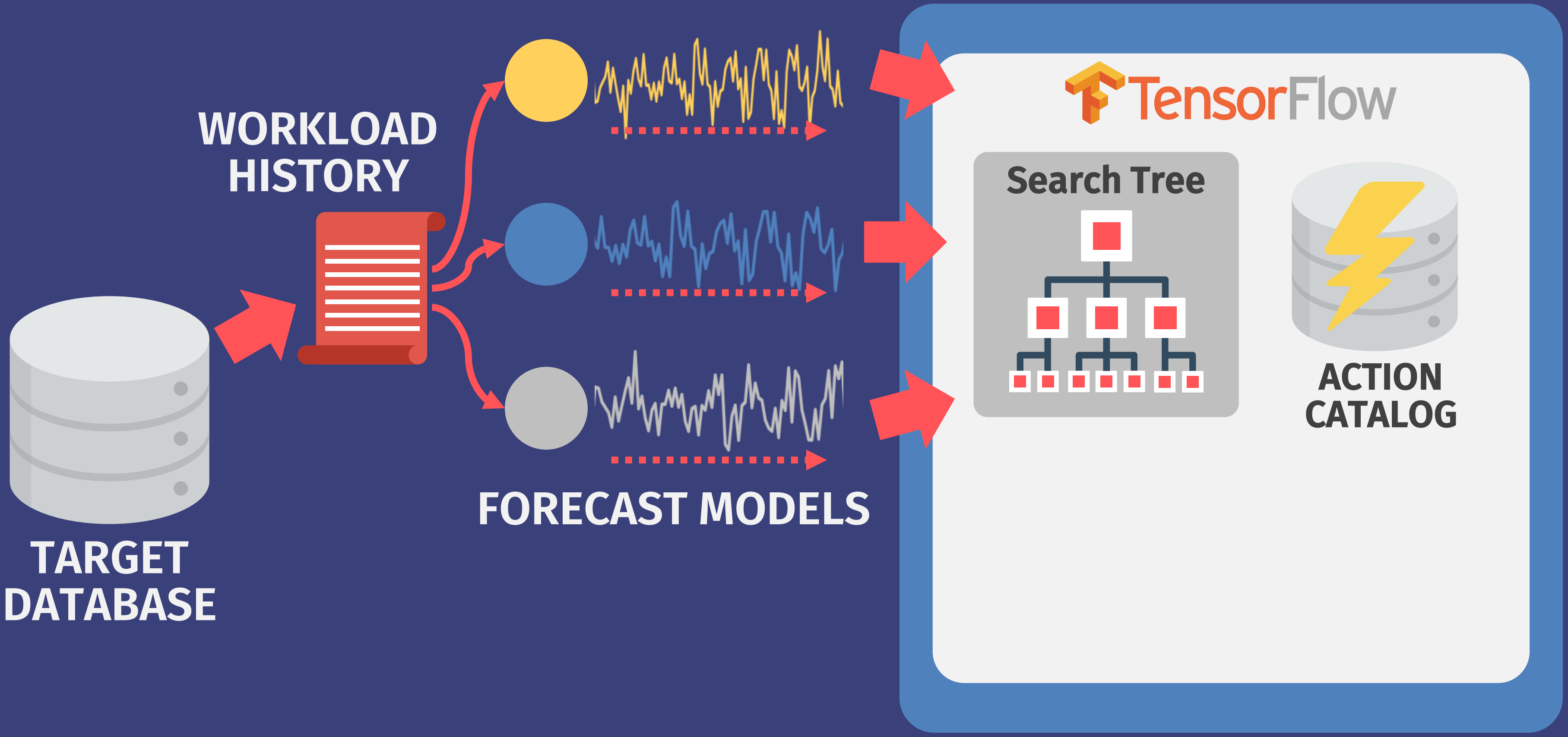
→ In-memory DBMS with integrated ML/RL framework.

→ Designed for autonomous operations.

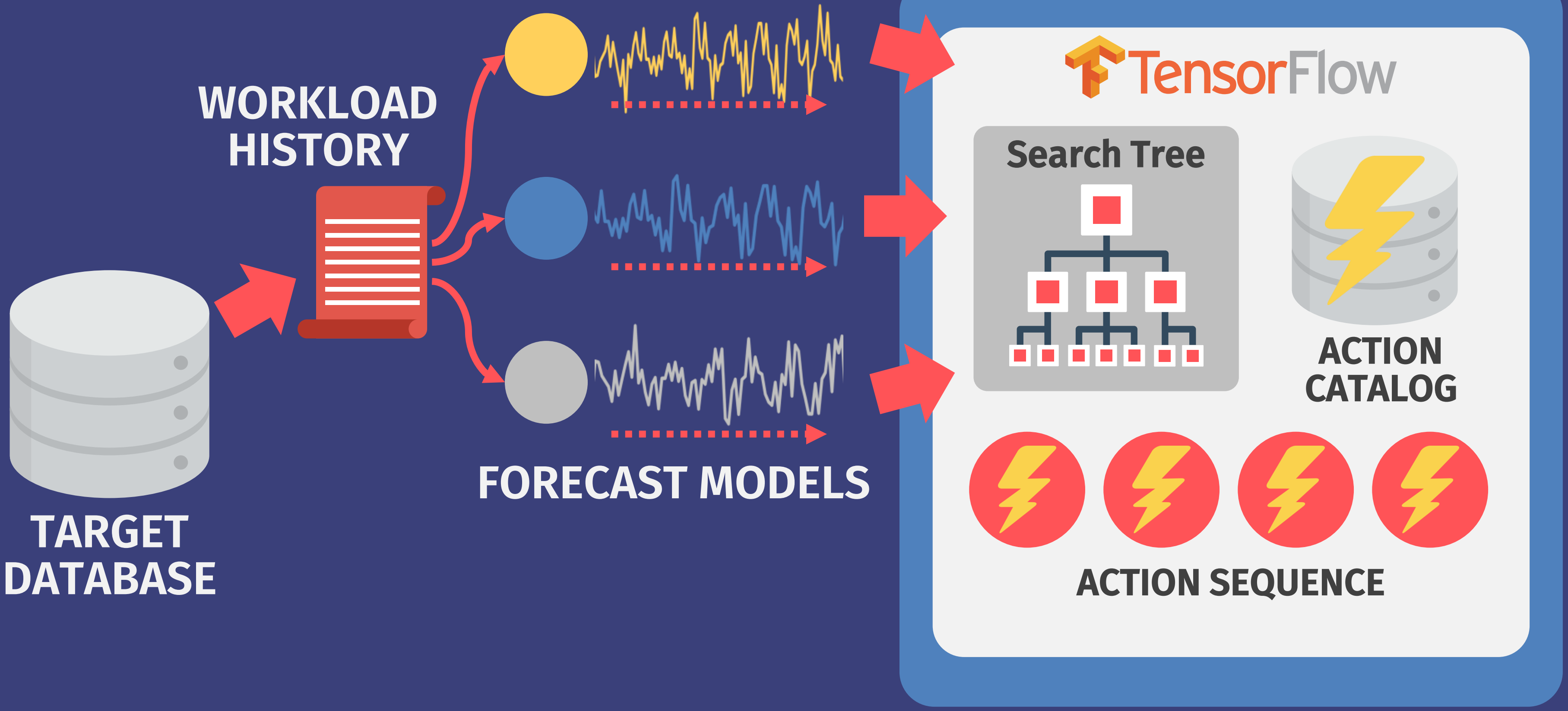

WORKLOAD HISTORY
TARGET DATABASE


WORKLOAD HISTORY
TARGET DATABASE
FORECAST MODELS

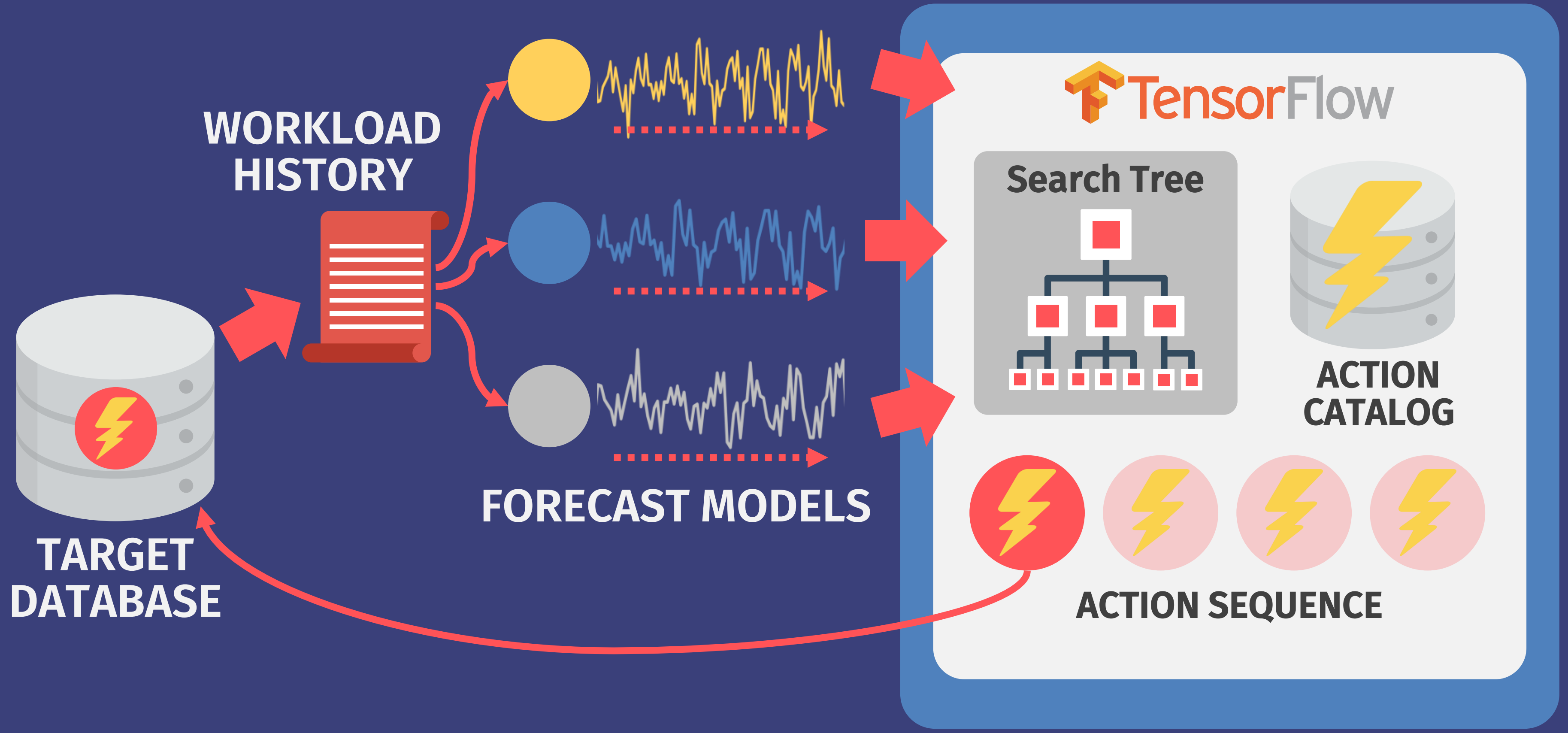
"THE BRAIN"


WORKLOAD HISTORY
TARGET DATABASE
FORECAST MODELS
TensorFlow
Search Tree
ACTION CATALOG

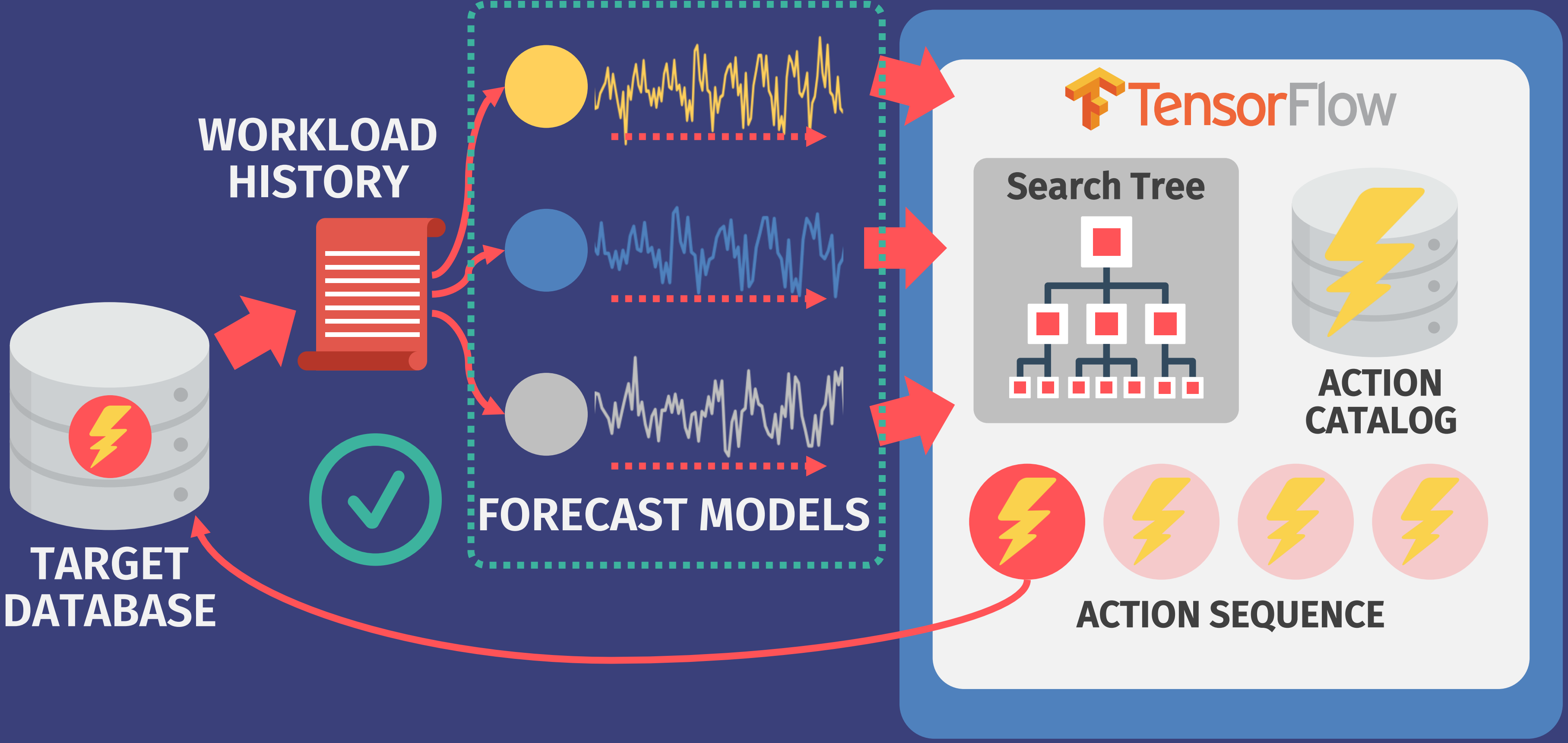
"THE BRAIN"


WORKLOAD
HISTORY
TARGET
DATABASE
FORECAST MODELS
TensorFlow
Search Tree
ACTION
CATALOG
ACTION SEQUENCE

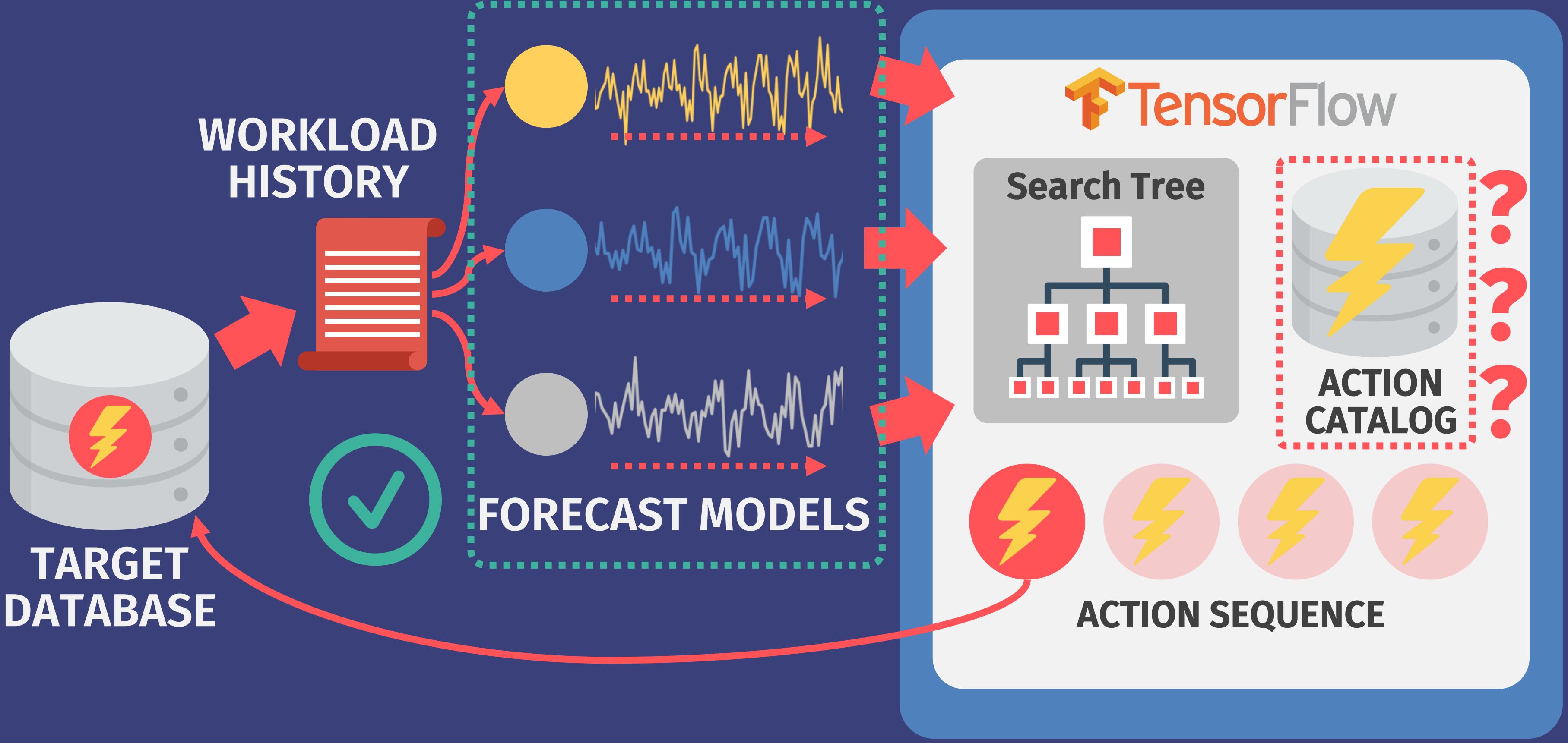
"THE BRAIN"


WORKLOAD
HISTORY
TARGET
DATABASE
FORECAST MODELS
TensorFlow
Search Tree
ACTION
CATALOG
ACTION SEQUENCE

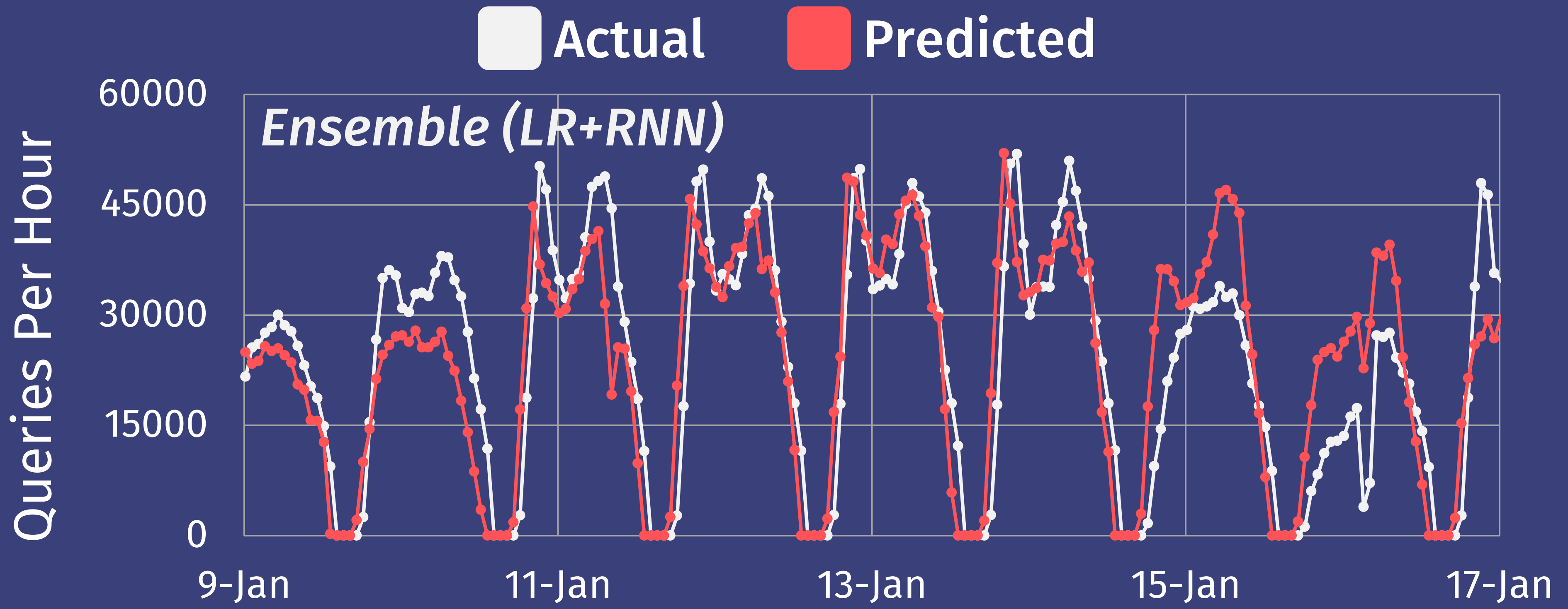
"THE BRAIN"


WORKLOAD HISTORY
FORECAST MODELS
TensorFlow
Search Tree
ACTION CATALOG
ACTION SEQUENCE
TARGET DATABASE

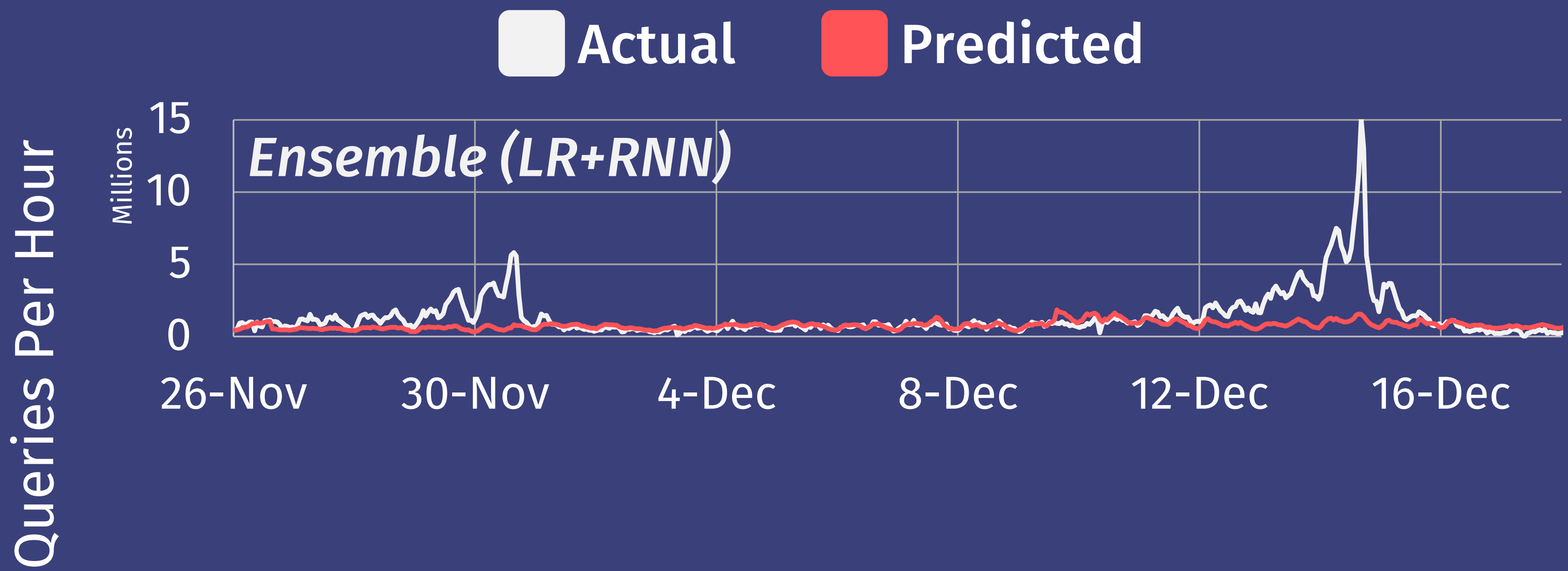
"THE BRAIN"

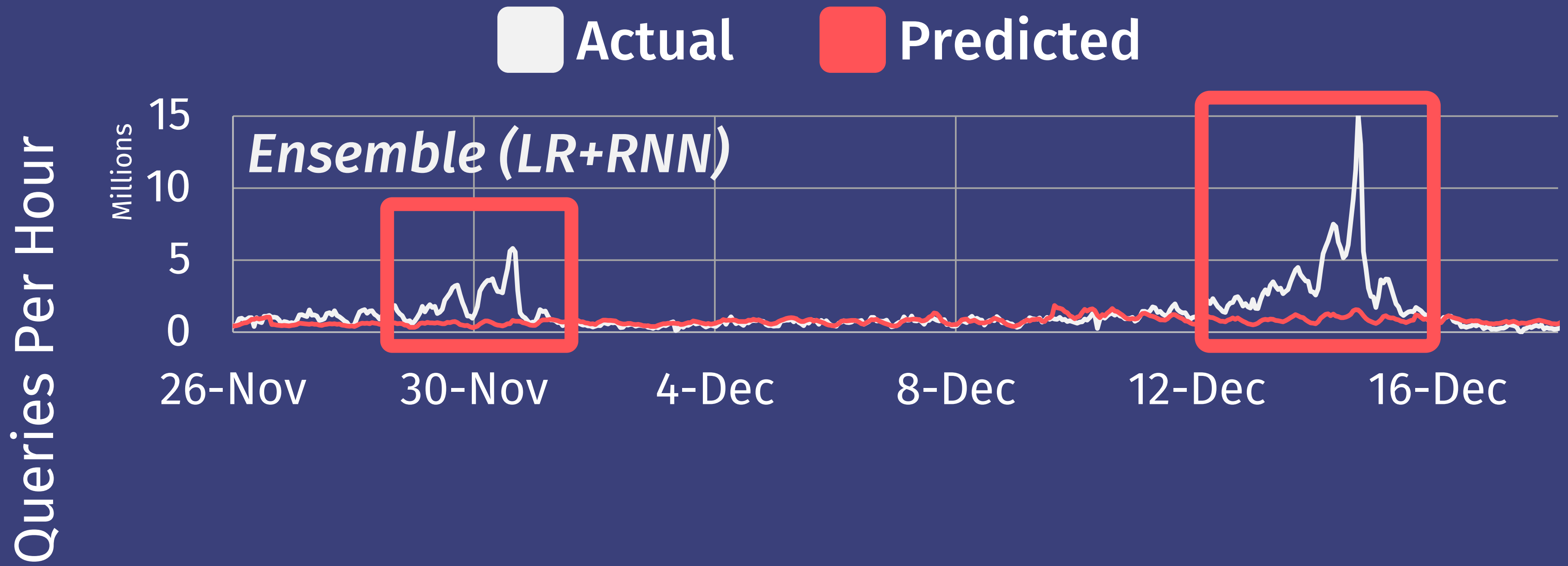

WORKLOAD HISTORY
TARGET DATABASE
FORECAST MODELS
TensorFlow
Search Tree
ACTION CATALOG
ACTION SEQUENCE

PELOTON
BUS TRACKING APP WITH ONE-HOUR HORIZON



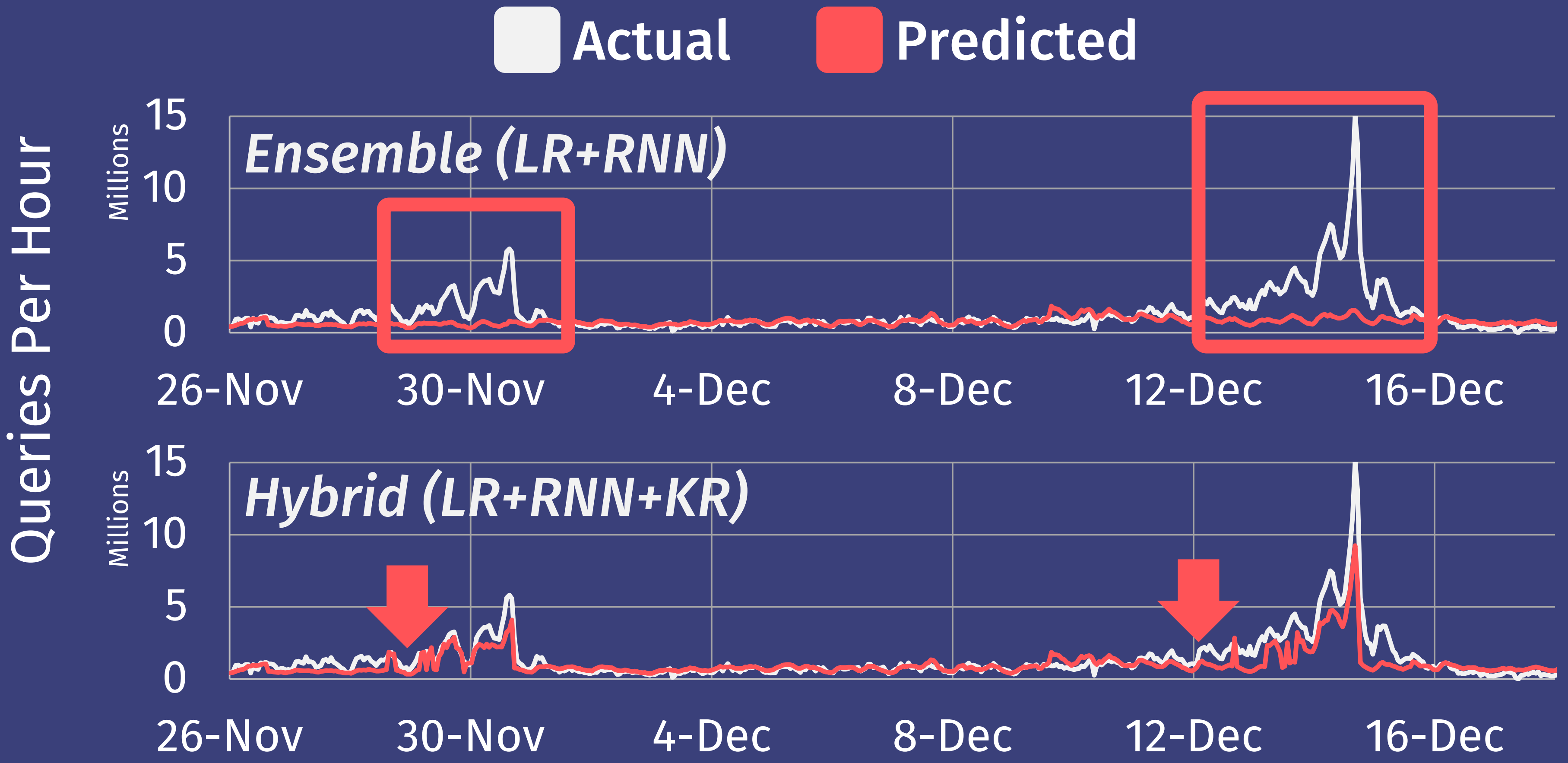
QUERY-BASED WORKLOAD FORECASTING FOR SELF-DRIVING DATABASE MANAGEMENT SYSTEM
SIGMOD 2018


Actual
Predicted
Ensemble (LR+RNN)
Queries Per Hour
Millions
15
10
5
0
26-Nov
30-Nov
4-Dec
8-Dec
12-Dec
16-Dec


Actual
Predicted
Ensemble (LR+RNN)
Queries Per Hour
Millions
15
10
5
0
26-Nov
30-Nov
4-Dec
8-Dec
12-Dec
16-Dec

PELOTON

ADMISSIONS APP WITH THREE-DAY HORIZON



Let's on check the demo...

Design Considerations for Autonomous Operation

Configuration Knobs

Internal Metrics

Action Engineering

Anything that requires a human value judgement should be marked as off-limits to autonomous components.

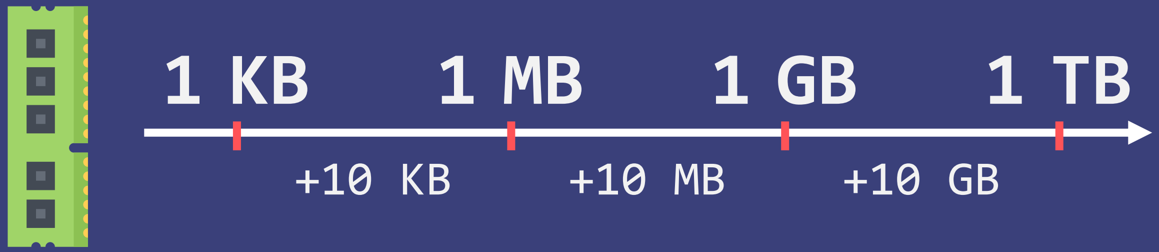
- *File Paths*
- *Network Addresses*
- *Durability / Isolation Levels*

The autonomous components need hints about how to change a knob

- *Min/max ranges.*
- *Separate knobs to enable/disable a feature.*
- *Non-uniform deltas.*

The autonomous components need hints about how to change a knob

- *Min/max ranges.*
- *Separate knobs to enable/disable a feature.*
- *Non-uniform deltas.*



CONFIGURATION KNOBS
HOW TO CHANGE

The autonomous com
about how to change

- *Min/max ranges.*
- *Separate knobs to e*
- *Non-uniform deltas.*

```
pavlo=> \d pg_settings;
     View "pg_catalog.pg_settings"
     Column      |  Type   | Modifiers
-----------------+---------+-----------
 name            | text    |
 setting         | text    |
 unit            | text    |
 category        | text    |
 short_desc      | text    |
 extra_desc      | text    |
 context         | text    |
 vartype         | text    |
 source          | text    |
 min_val         | text    |
 max_val         | text    |
 enumvals        | text[]  |
 boot_val        | text    |
 reset_val       | text    |
 sourcefile      | text    |
 sourceline      | integer |
 pending_restart | boolean |
```

Indicate which knobs are constrained by hardware resources.

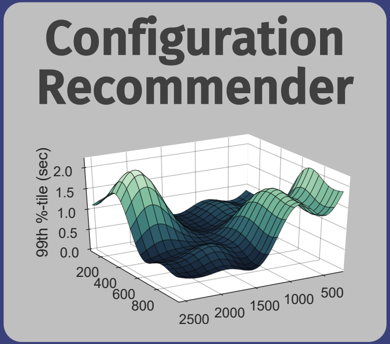
- *The sum of all buffers cannot exceed the total amount of available memory.*

The problem is that sometimes it makes sense to overprovision.

Expose DBMS's hardware capabilities:

– *CPU, Memory, Disk, Network*

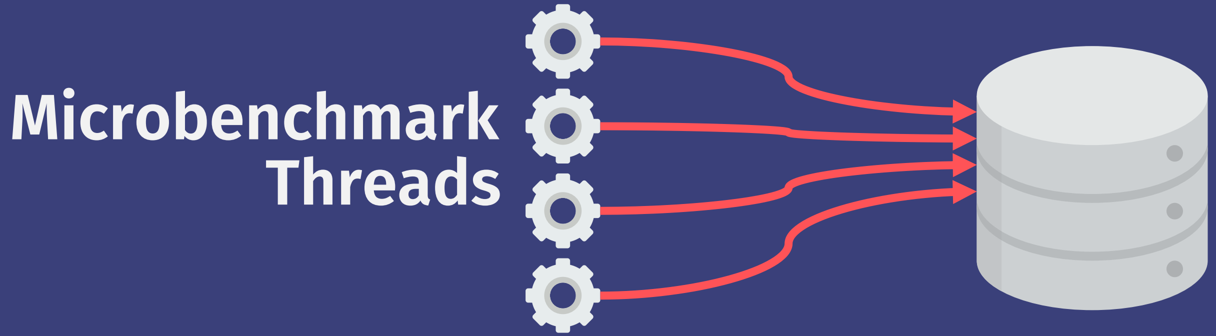
```
pavlo=# SELECT * FROM INFORMATION_SCHEMA.cpuinfo ;
 id |       model        |   mhz    |  cache  | bogomips
----+--------------------+----------+---------+----------
  0 | Intel Xeon E7-4830 | 3472.718 | 4096 KB |  5615.84
  1 | Intel Xeon E7-4830 | 3472.718 | 4096 KB |  5615.84
  2 | Intel Xeon E7-4830 | 3472.718 | 4096 KB |  5615.84
  3 | Intel Xeon E7-4830 | 3472.718 | 4096 KB |  5615.84
  4 | Intel Xeon E7-4830 | 3472.718 | 4096 KB |  5615.84
  5 | Intel Xeon E7-4830 | 3472.718 | 4096 KB |  5615.84
  6 | Intel Xeon E7-4830 | 3472.718 | 4096 KB |  5615.84
  7 | Intel Xeon E7-4830 | 3472.718 | 4096 KB |  5615.84
```



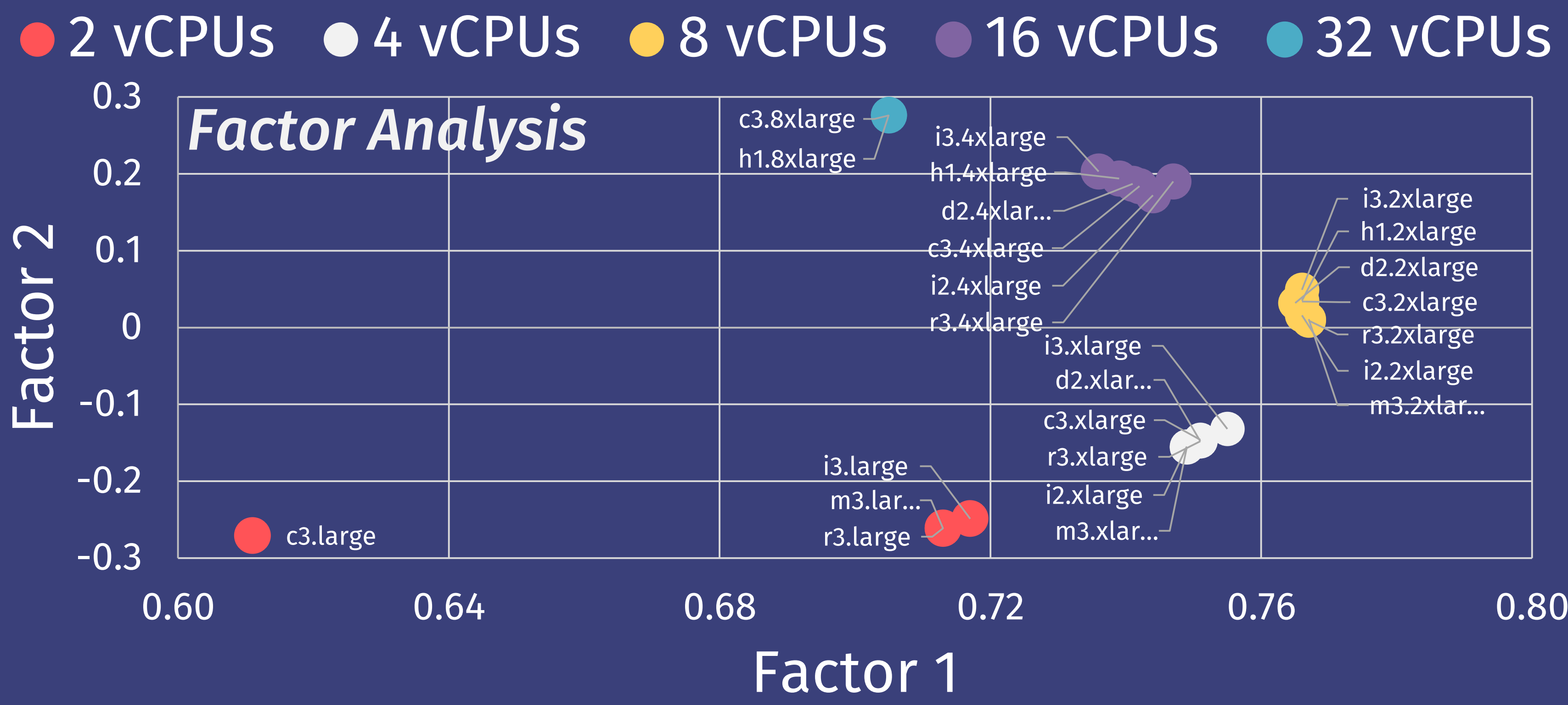
Expose DBMS's hardware capabilities:

– *CPU, Memory, Disk, Network*

Otherwise you have to come up with clever ways to approximate this…



INTERNAL METRICS
HARDWARE MICROBENCHMARKS



If the DBMS has sub-components that are tunable, then it must expose separate metrics for those components.

<u>Bad</u> Example: RocksDB

RocksDB Column Family Knobs

```
rocksdb_override_cf_options=\
     cf_link_pk={prefix_extractor=capped:20}
```

Column Family Metrics

```
mysql> SELECT * FROM INFORMATION_SCHEMA.ROCKSDB_CFSTAT;
+----------+-------------------------------------+-----------+
|  CF_NAME | METRIC_NAME                         | VALUE     |
+----------+-------------------------------------+-----------+
|  default | COMPACTION_PENDING                  | 1         |
|  default | CUR_SIZE_ACTIVE_MEM_TABLE           | 21672     |
|  default | CUR_SIZE_ALL_MEM_TABLES             | 21672     |
|  default | MEM_TABLE_FLUSH_PENDING             | 0         |
|  default | NON_BLOCK_CACHE_SST_MEM_USAGE       | 0         |
|  default | NUM_ENTRIES_ACTIVE_MEM_TABLE        | 18        |
|  default | NUM_ENTRIES_IMM_MEM_TABLES          | 0         |
|  default | NUM_IMMUTABLE_MEM_TABLE             | 0         |
|  default | NUM_LIVE_VERSIONS                   | 2         |
+----------+-------------------------------------+-----------+
```

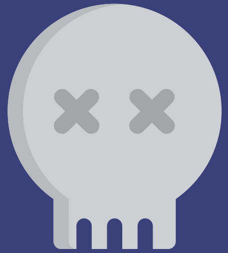
Missing:
Reads
Writes

RocksDB Column Family Knobs

```
rocksdb_override_cf_options=\
     cf_link_pk={prefix_extractor=capped:20}
```

Global Metrics

```
mysql> SHOW GLOBAL STATUS;
+-----------------------------------+-----------+
| METRIC_NAME                       | VALUE     |
+-----------------------------------+-----------+
| ABORTED_CLIENTS                   | 0         |
•••
| ROCKSDB_BLOCK_CACHE_BYTES_READ    | 295700537 |
| ROCKSDB_BLOCK_CACHE_BYTES_WRITE   | 709562185 |
| ROCKSDB_BLOCK_CACHE_DATA_HIT      | 64184     |
| ROCKSDB_BLOCK_CACHE_DATA_MISS     | 1001083   |
| ROCKSDB_BYTES_READ                | 5573794   |
| ROCKSDB_BYTES_WRITTEN             | 5817440   |
| ROCKSDB_FLUSH_WRITE_BYTES         | 2906847   |
•••
| UPTIME_SINCE_FLUSH_STATUS         | 5996      |
+-----------------------------------+-----------+
```



Aggregated Metrics

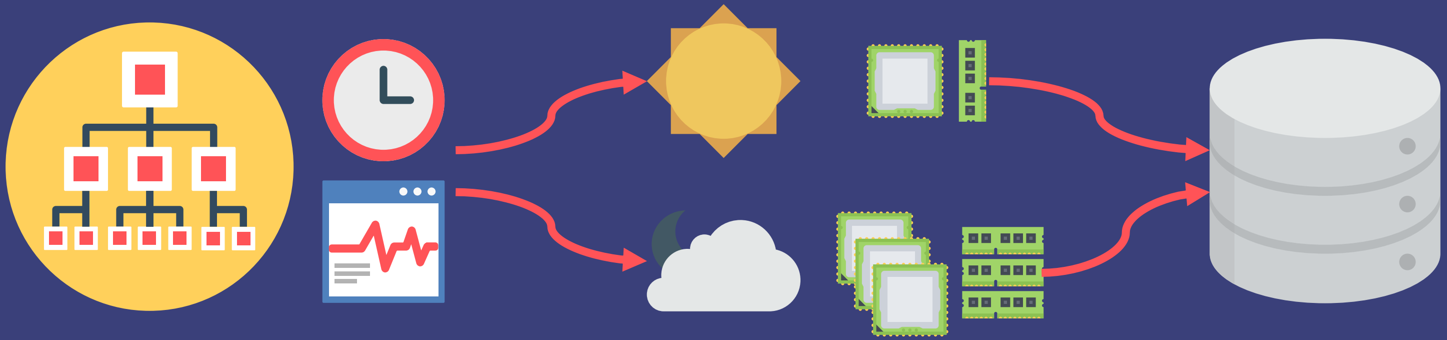
No action should ever require the DBMS to restart in order for it to take affect.

The commercial systems are much better than this than the open-source systems.

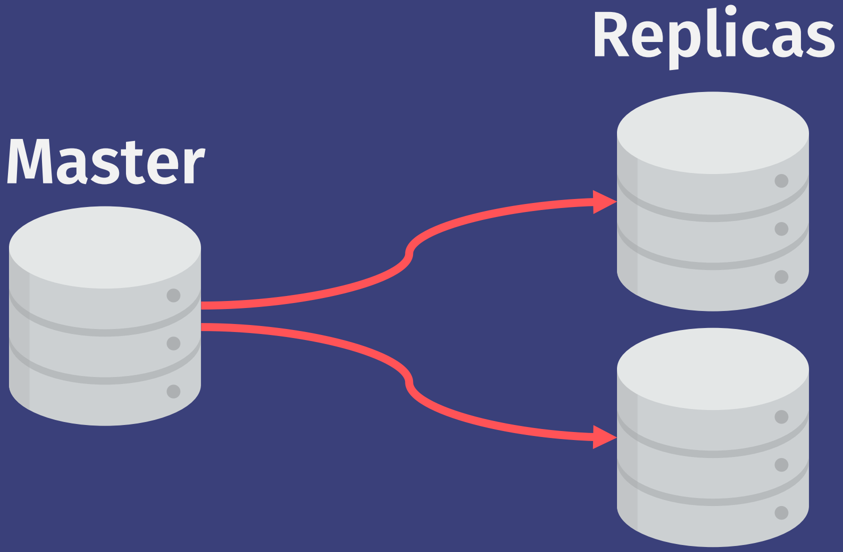
Provide a notification callback to indicate when an action starts and when it completes.

Harder for changes that can be used before the action completes.

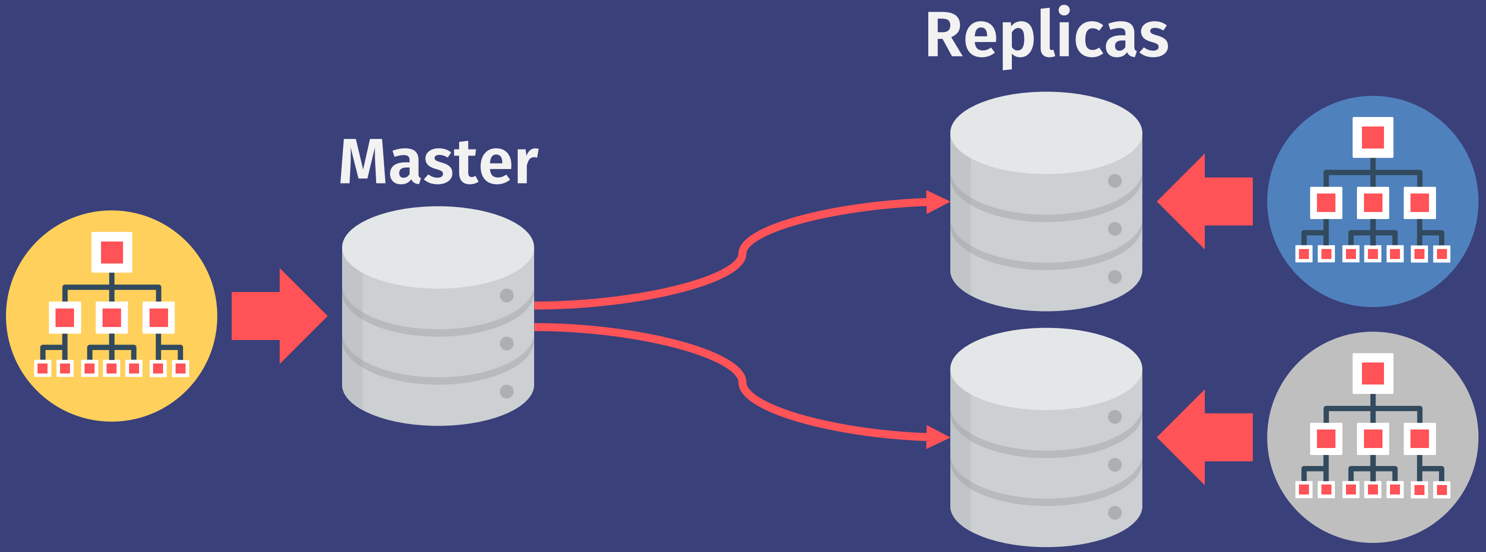
Support executing the same action with different resource usage levels.



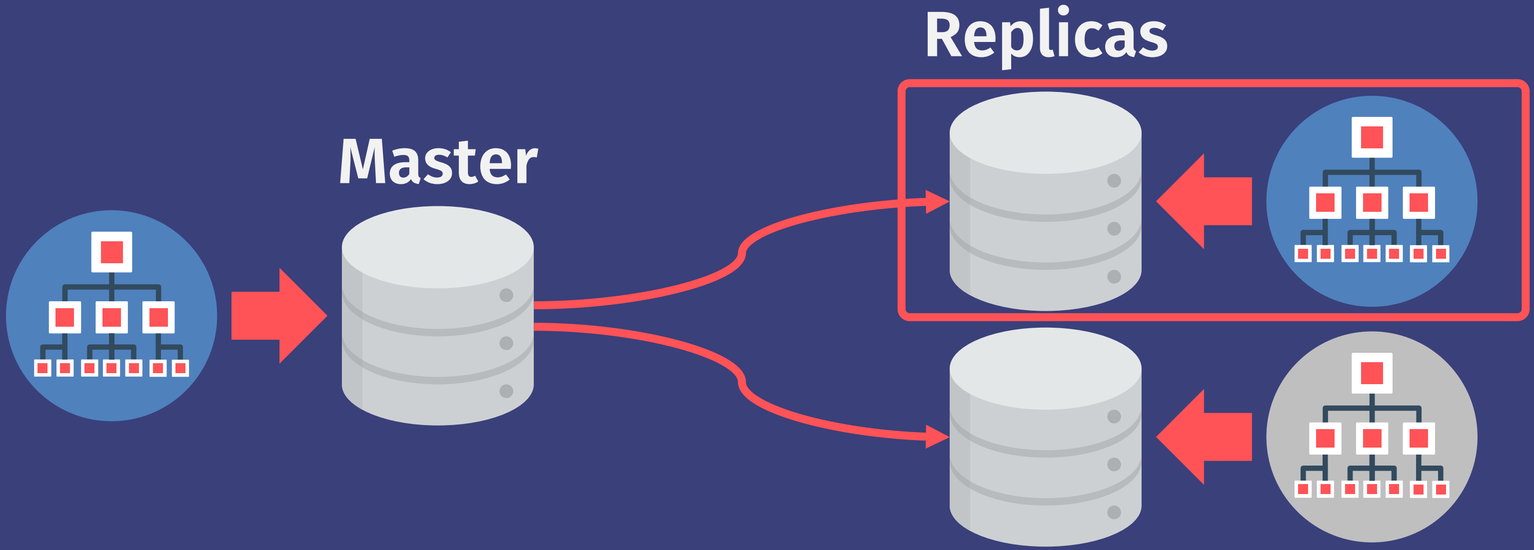
Allow replica configurations to diverge from each other.



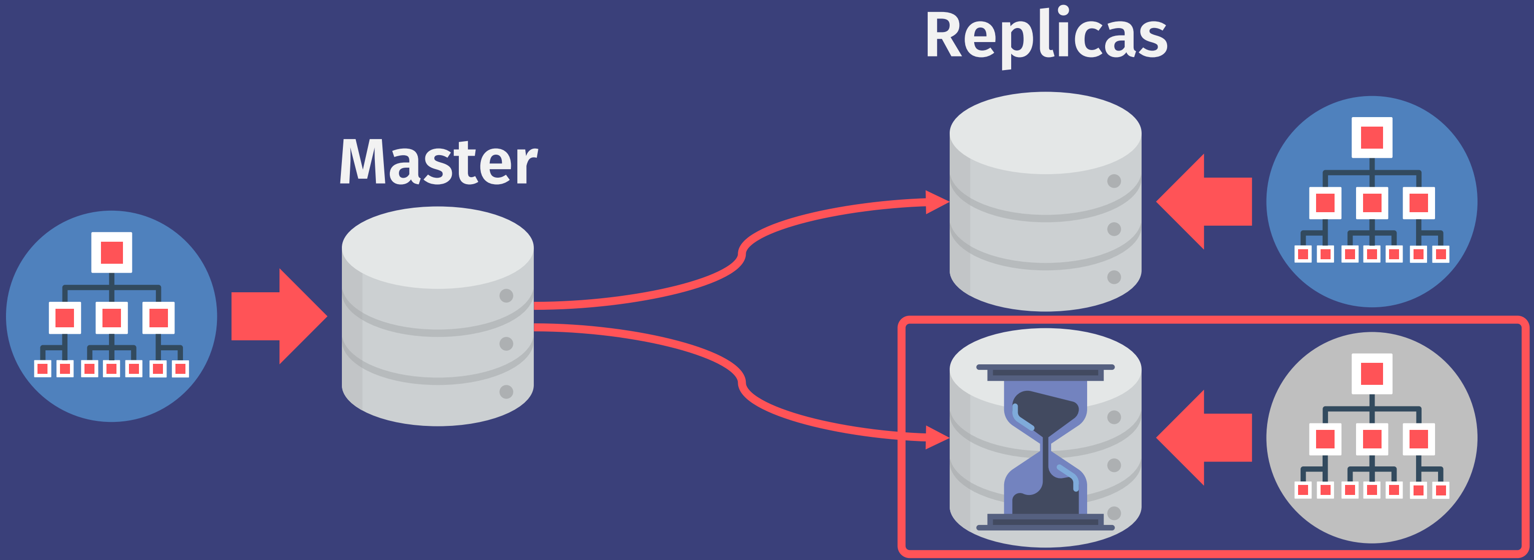
Allow replica configurations to diverge from each other.



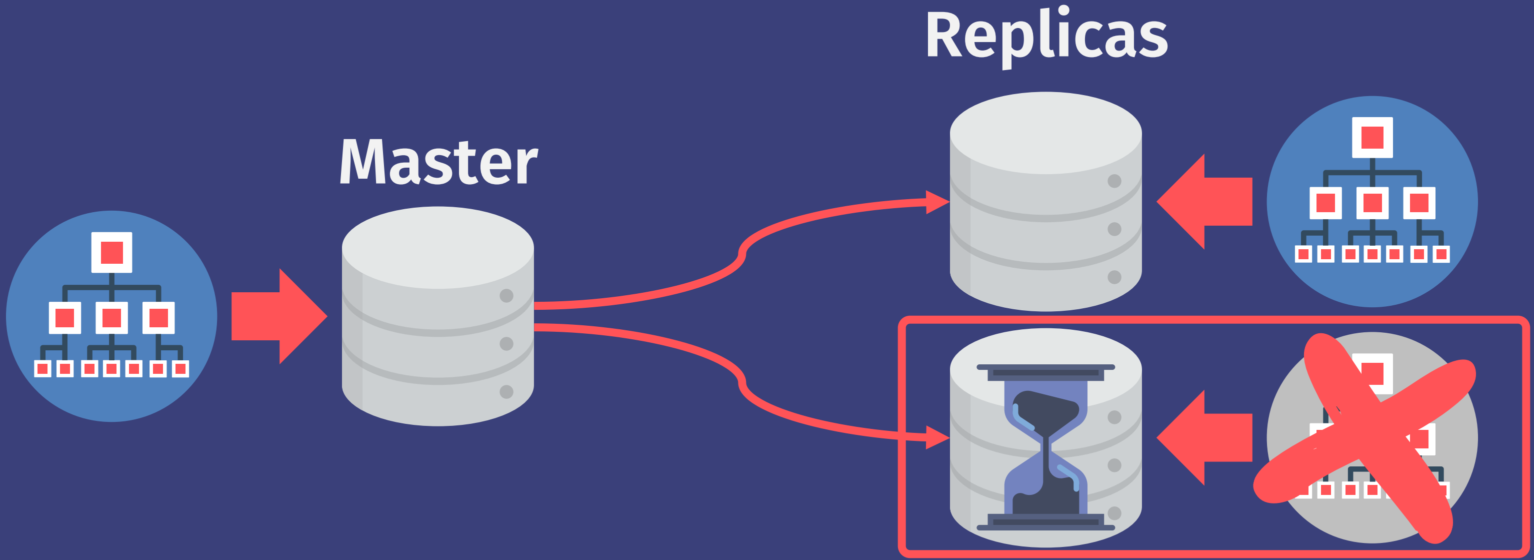
Allow replica configurations to diverge from each other.



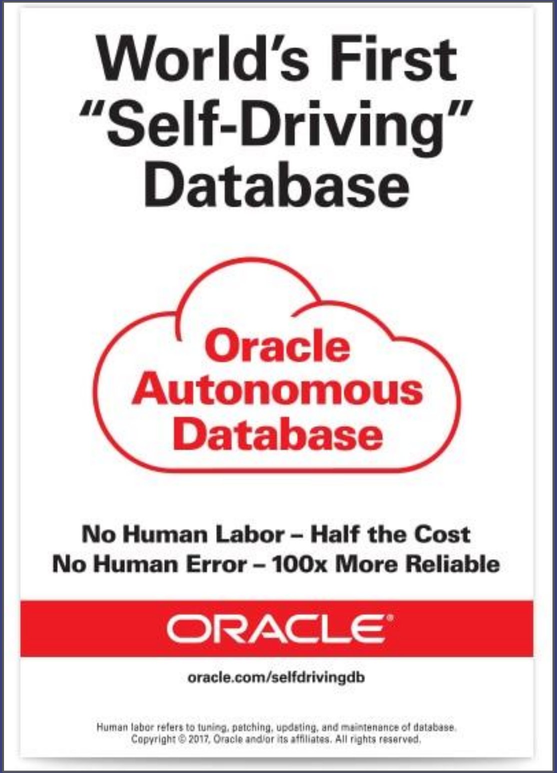
Allow replica configurations to diverge from each other.



Allow replica configurations to diverge from each other.



What About Oracle's Self-Driving DBMS?



September 2017

Self-Driving Database Management Systems

Andrew Pavlo, Gustavo Angulo, Joy Arulraj, Haibin Lin, Jiexi Lin, Lin Ma, Prashanth Menon
Todd C. Mowry, Matthew Perron, Ian Quah, Siddharth Santurkar, Anthony Tomasic
Skye Toor, Dana Van Aken, Ziqi Wang, Yingjun Wu*, Ran Xian, Tieying Zhang
Carnegie Mellon University, *National University of Singapore

ABSTRACT

In the last two decades, both researchers and vendors have built advisory tools to assist database administrators (DBAs) in various aspects of system tuning and physical design. Most of this previous work, however, is incomplete because they still require humans to make the final decisions about any changes to the database and are reactionary measures that fix problems after they occur.

What is needed for a truly "self-driving" database management system (DBMS) is a new architecture that is designed for autonomous operation. This is different than earlier attempts because all aspects of the system are controlled by an integrated planning component that not only optimizes the system for the current workload, but also predicts future workload trends so that the system can prepare itself accordingly. With this, the DBMS can support all of the previous tuning techniques without requiring a human to determine the right way and proper time to deploy them. It also enables new optimizations that are important for modern high-performance DBMSs, but which are not possible today because the complexity of managing these systems has surpassed the abilities of human experts.

This paper presents the architecture of Peloton, the first self-driving DBMS. Peloton's autonomic capabilities are now possible due to algorithmic advancements in deep learning, as well as improvements in hardware and adaptive database architectures.

1. INTRODUCTION

The idea of using a DBMS to remove the burden of data management was one of the original selling points of the relational model and declarative query languages from the 1970s [31]. With this approach, a developer only writes a query that specifies what data they want to access. The DBMS then finds the most efficient way to store and retrieve data, and to safely interleave operations.

Over four decades later, DBMSs are now the critical part of every data-intensive application in all facets of society, business, and science. These systems are also more complicated now with a long and growing list of functionalities. But using existing automated tuning tools is an onerous task, as they require laborious preparation of workload samples, spare hardware to test proposed updates, and above all else intuition into the DBMS's internals. If the DBMS could do these things automatically, then it would remove many of the complications and costs involved with deploying a database [40].

Much of the previous work on self-tuning systems is focused on standalone tools that target only a single aspect of the database. For example, some tools are able to choose the best logical or physical design of a database [16], such as indexes [30, 17, 58], partitioning schemes [6, 44], data organization [7], or materialized views [5]. Other tools are able to select the tuning parameters for an application [56, 22]. Most of these tools operate in the same way: the DBA provides it with a sample database and workload trace that guides a search process to find an optimal or near-optimal configuration. All of the major DBMS vendors' tools, including Oracle [23, 38], Microsoft [16, 42], and IBM [55, 57], operate in this manner. There is a recent push for integrated components that support adaptive architectures [36], but these again only focus on solving one problem. Likewise, cloud-based systems employ dynamic resource allocation at the service-level [20], but do not tune individual databases.

All of these are insufficient for a completely autonomous system because they are (1) external to the DBMS, (2) reactionary, or (3) not able to take a holistic view that considers more than one problem at a time. That is, they observe the DBMS's behavior from outside of the system and advise the DBA on how to make corrections to fix only one aspect of the problem after it occurs. The tuning tools assume that the human operating them is knowledgeable enough to update the DBMS during a time window when it will have the least impact on applications. The database landscape, however, has changed significantly in the last decade and one cannot assume that a DBMS is deployed by an expert that understands the intricacies of database optimization. But even if these tools were automated such that they could deploy the optimizations on their own, existing DBMS architectures are not designed to support major changes without stressing the system further nor are they able to adapt in anticipation of future bottlenecks.

In this paper, we make the case that self-driving database systems are now achievable. We begin by discussing the key challenges with such a system. We then present the architecture of **Peloton** [1], the first DBMS that is designed for autonomous operation. We conclude with some initial results on using Peloton's integrated deep learning framework for workload forecasting and action deployment.

2. PROBLEM OVERVIEW

The first challenge in a self-driving DBMS is to understand an application's workload. The most basic level is to characterize queries as being for either an OLTP or OLAP application [26]. If the DBMS identifies which of these two workload classes the application belongs to, then it can make decisions about how to optimize the database. For example, if it is OLTP, then the DBMS should store tuples in a row-oriented layout that is optimized for writes. If it is OLAP, then the DBMS should use a column-oriented

8th Biennial Conference on Innovative Data Systems Research (CIDR '17) January 8-11, 2017 , Chaminade, California, USA.

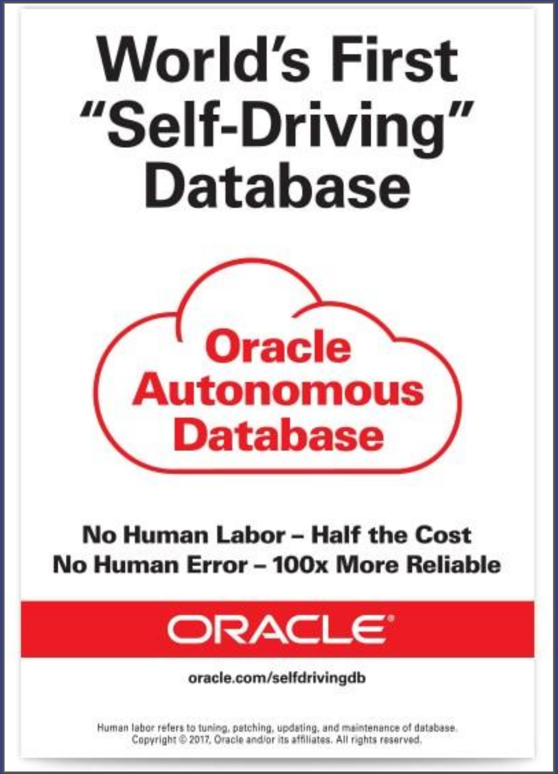
January 2017

Automatic Indexing

Automatic Recovery

Automatic Scaling

Automatic Query Tuning



September 2017

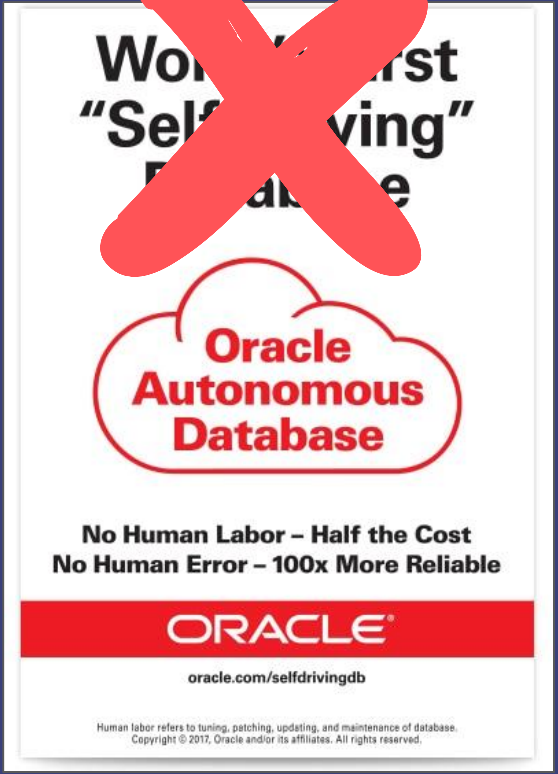
Automatic Indexing

Automatic Recovery

Automatic Scaling

Problem #2
Reactionary Measures

Automatic Query Tuning



September 2017

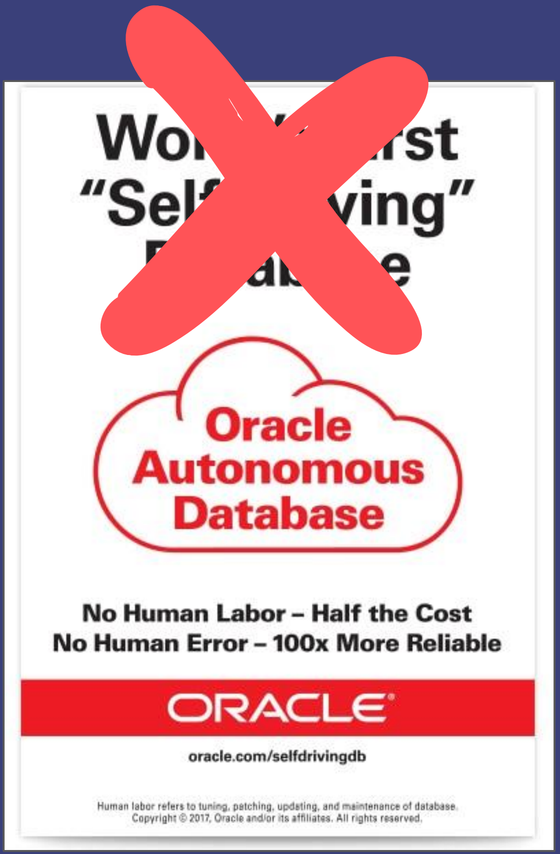
Automatic Indexing

Automatic Recovery

Automatic Scaling

Problem #2
Reactionary Measures

Automatic Query Tuning



September 2017

Automatic Indexing

Automatic Recovery

Automatic Scaling

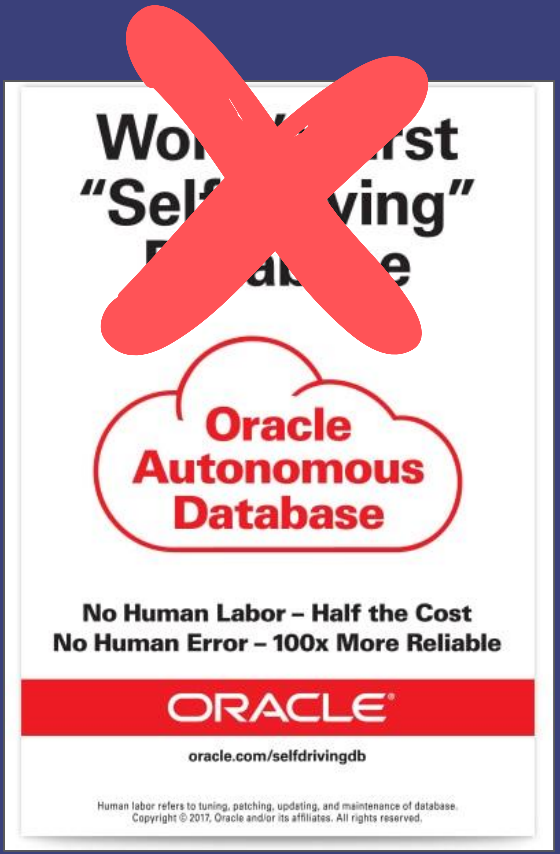
Problem #2
Reactionary Measures

Automatic Query Tuning

Microsoft

IBM DB2

INGRES



September 2017

True autonomous DBMSs are achievable in the next decade.

You should think about how each new feature can be controlled by a machine.

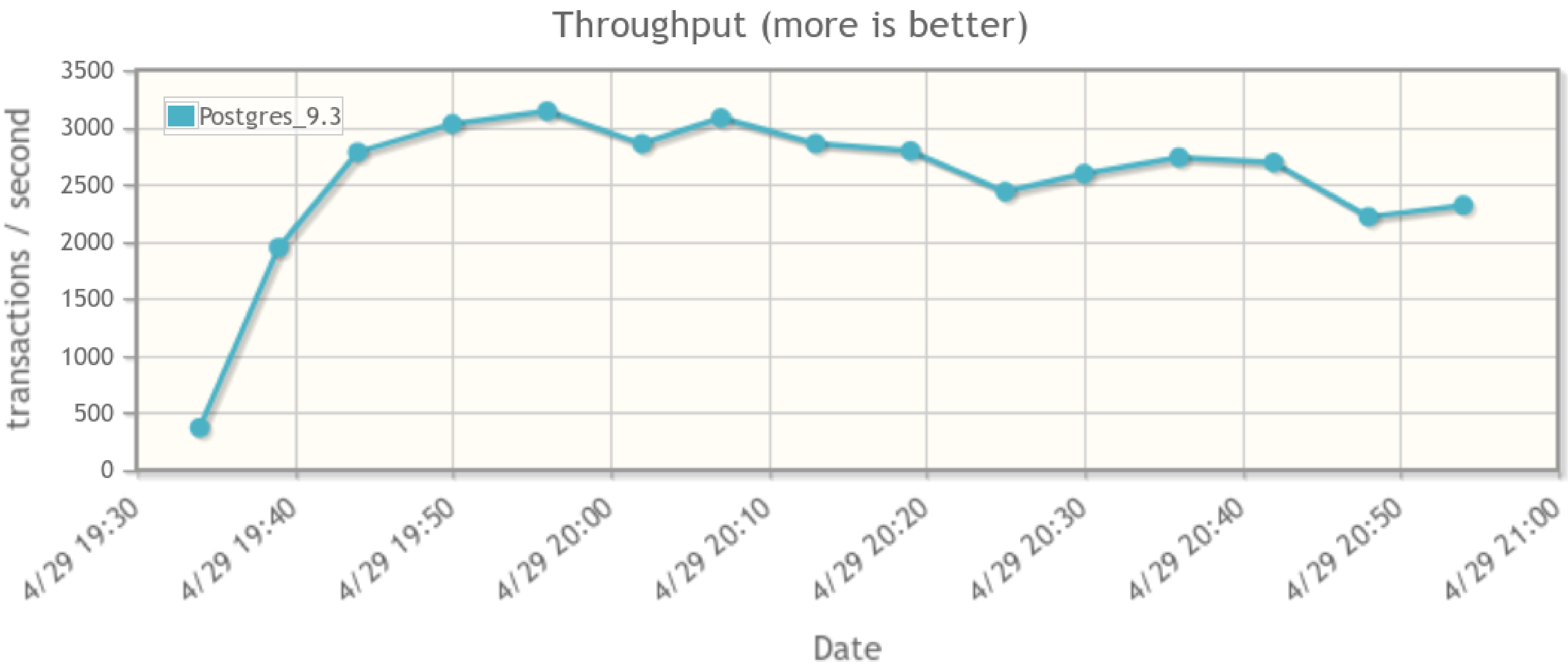
Demo Results

END

@andy_pavlo

Show the last 100 results

Eq



l

l_fsync