Crowdsourced Databases: Query Processing with People

Adam Marcus, Eugene Wu,
David Karger, Sam Madden, Rob Miller

MIT CSAIL

Crowdsourced Databases: Query Processing with People

How to Crowdsource the Introduction of Your Talk to Other Research Groups

CIDR Deadline

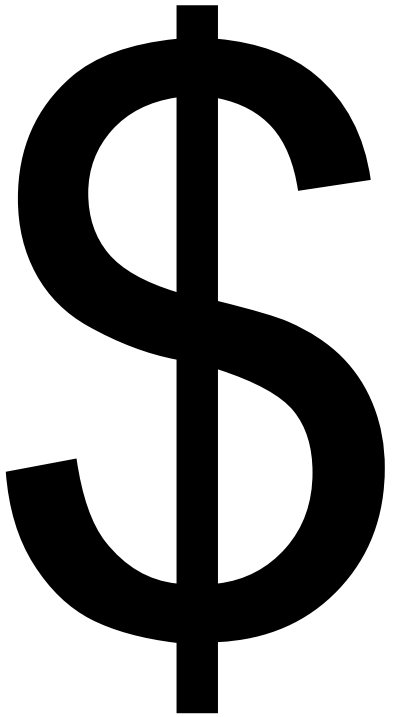
Good ideas in Databases

Jane, John

Abstract

Traditional databases ***fill in here***.
This leads to ***fill in here***.
We propose ***find a good name***,
which ***what does it do?***.

OR



CIDR 2011

January 9-12, 2011

Conference Organization

Conference Chairs

- **Anastasia Ailamaki**, EPFL
- **Michael Franklin**, UC Berkeley
- **Joe Hellerstein**, UC Berkeley

CIDR 2011

January 9-12, 2011

- Home
- Call For Papers
- Organization
- Program

Conference Organization

Conference Chairs

- **Anastasia Ailamaki**, EPFL
- **Michael Franklin**, UC Berkeley
- **Joe Hellerstein**, UC Berkeley

D2 | dna@cs.yale.edu

	A	B	C	D
1	Name	Affiliation	Phone	Email
2	Daniel Abadi	Yale	203.436.1265	dna@cs.yale.edu
3	Anastasia Ailamaki	EPFL		
4	Gustavo Alonso	ETH Zurich		
5	Sihem Amer Yahia	Yahoo! Research		
6	Magdalena Balazinska	Washington		
7	Michael Cafarella	Michigan		

CIDR 2011

January 9-12, 2011

- Home
- Call For Papers
- Organization
- Program

Conference Organization

Conference Chairs

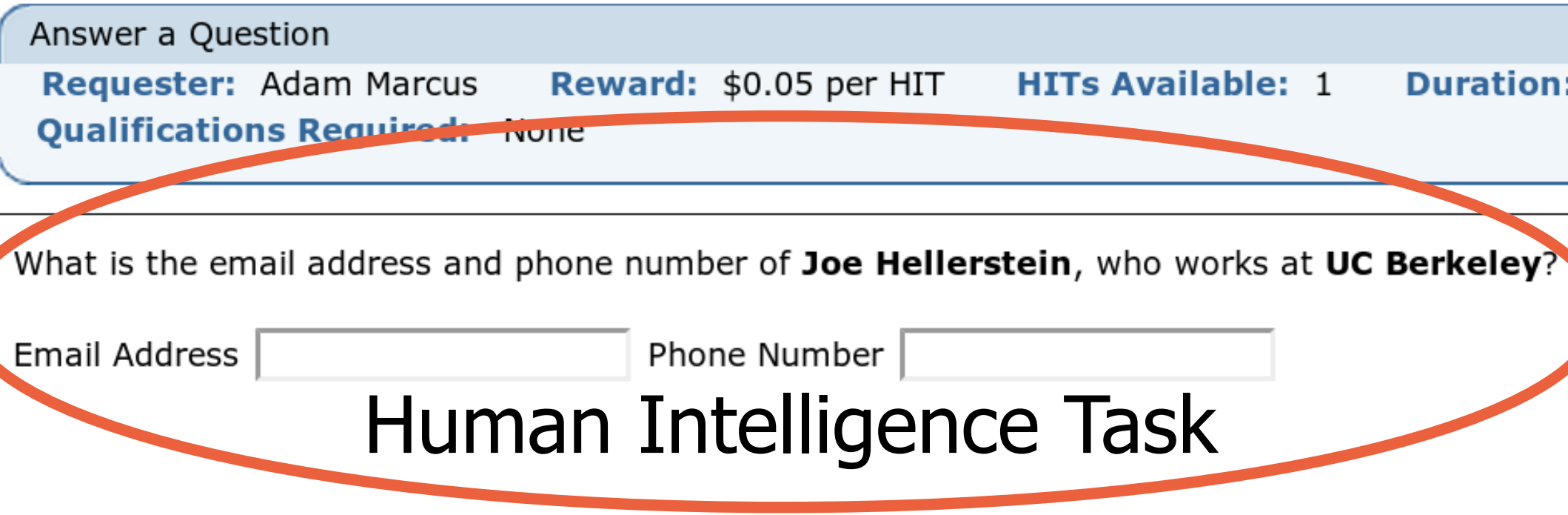
- **Anastasia Ailamaki**, EPFL
- **Michael Franklin**, UC Berkeley
- **Joe Hellerstein**, UC Berkeley

D2 | dna@cs.yale.edu

	A	B	C	D
1	Name	Affiliation	Phone	Email
2	Daniel Abadi	Yale	203.436.1265	dna@cs.yale.edu
3	Anastasia Ailamaki	EPFL		
4	Gustavo Alonso	ETH Zurich		
5	Sihem Amer Yahia	Yahoo! Research		
6	Magdalena Balazinska	Washington		
7	Michael Cafarella	Michigan		


amazon mechanical turk
beta
Artificial Artificial Intelligence

Turker Interface



(HIT)

Uses of Human Computation

- Data cleaning/integration (ProPublica)
- Finding missing people (Haiti, Fossett, Gray)
- Translation/Transcription (SpeakerText)
- Word Processing (Soylent)
- Outsourced insurance claims processing
- Data journalism (Guardian)

Challenges in Human Computation

- Workers are not silicon drones
- Worker latency is minutes, hours
- Optimization parameters (price, accuracy) and model are different

Challenges in Human Computation

- Workers are not silicon drones
- Worker latency is minutes, hours
- Optimization parameters (price, accuracy) and model are different

These problems are currently addressed in an ad-hoc way

Qurk is a declarative workflow management system that allows human computation over relational data

Qurk is a declarative workflow management system that allows human computation over relational data

(a system which throws humans in the direct path of query execution)

Data Model/Query Model

- UDFs that compile to HTML forms
- List types
  - Make relational with Pig-style FLATTEN operator
  - Collapse to single values with UDAs

TASK pcInfo(name, affiliation)

Text: “What is the email address and phone number of %s, who works at %s”, name, affiliation

Response: Form((“Email Address”,String), (“Phone Number”,String))

```
SELECT name, FLATTEN(pcInfo(name, affiliation))
FROM pc
INTO temporary Q1
```

```
SELECT name, FLATTEN(pcInfo(name, affiliation))
FROM pc
INTO temporary Q1
```

```
      (h@berkeley, 123-456-7890)
Joe,((h@berkeley, (999)4445555) )
      (bad @ response,1234567890)
```


```
Joe, h@berkeley, 123-456-7890
Joe, h@berkeley, (999)4445555)
Joe, bad @ response,1234567890
```

```
SELECT name, majorityVote(normalize(phone)),
                    majorityVote(normalize(email)),
FROM Q1
GROUP BY name
```

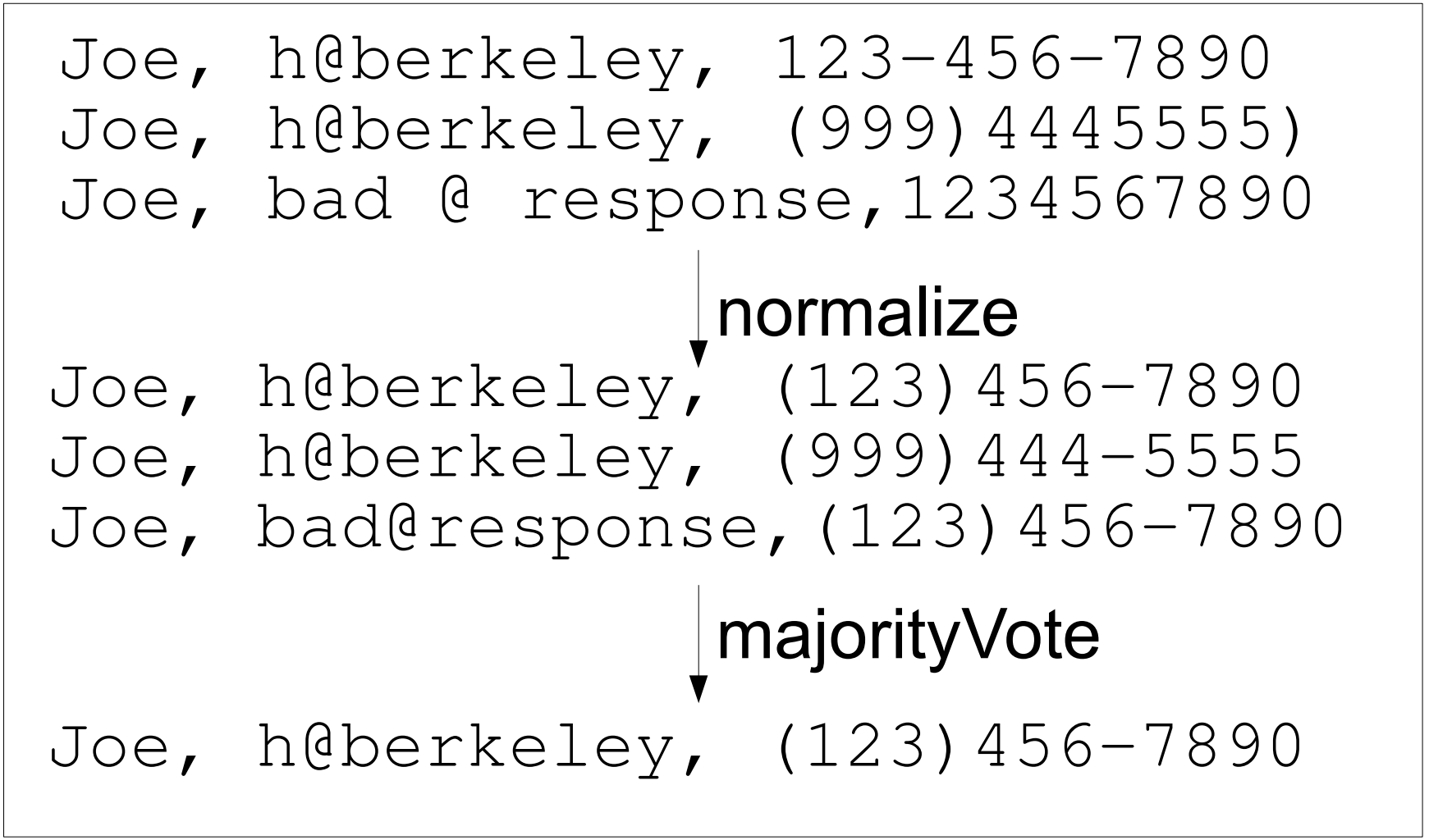
```
SELECT name, majorityVote(normalize(phone)),
                    majorityVote(normalize(email)),
FROM Q1
GROUP BY name
```

```
Joe, h@berkeley, 123-456-7890
Joe, h@berkeley, (999)4445555)
Joe, bad @ response,1234567890
```

normalize

```
Joe, h@berkeley, (123)456-7890
Joe, h@berkeley, (999)444-5555
Joe, bad@response,(123)456-7890
```

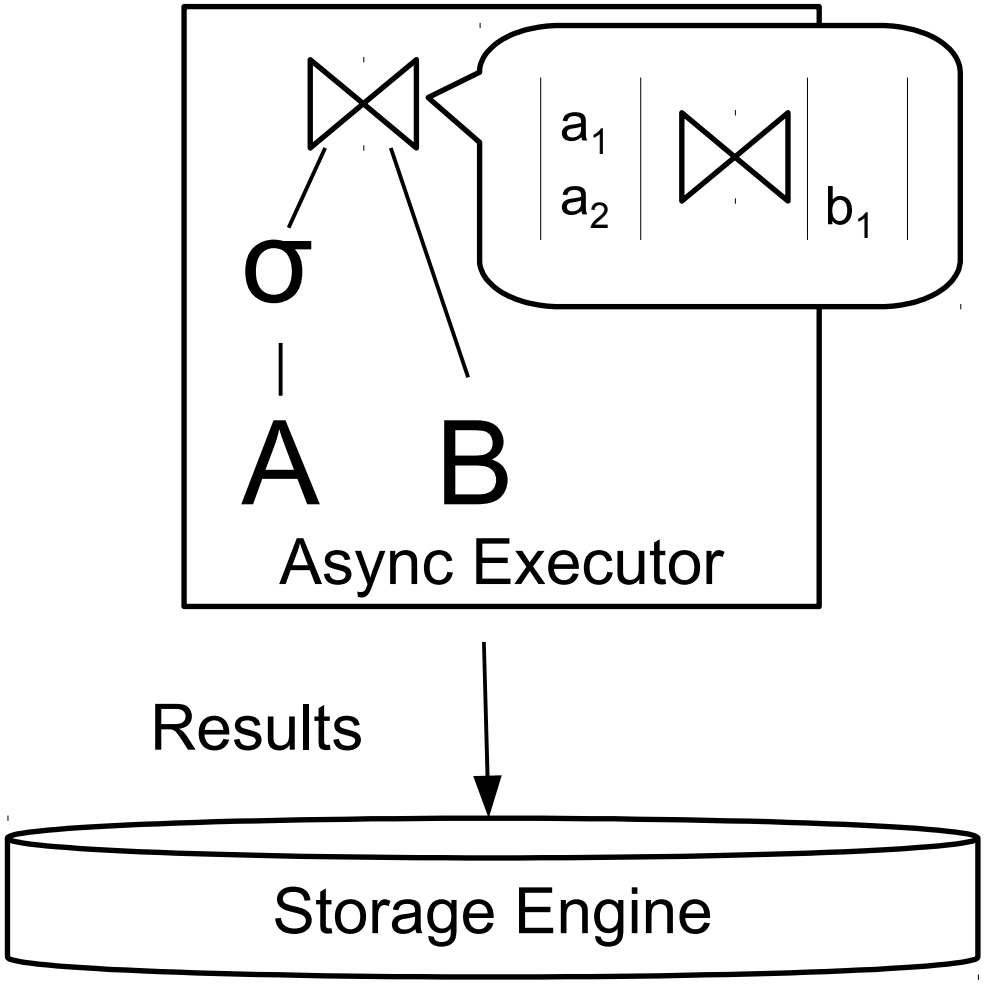
```
SELECT name, majorityVote(normalize(phone)),
                majorityVote(normalize(email)),
FROM Q1
GROUP BY name
```

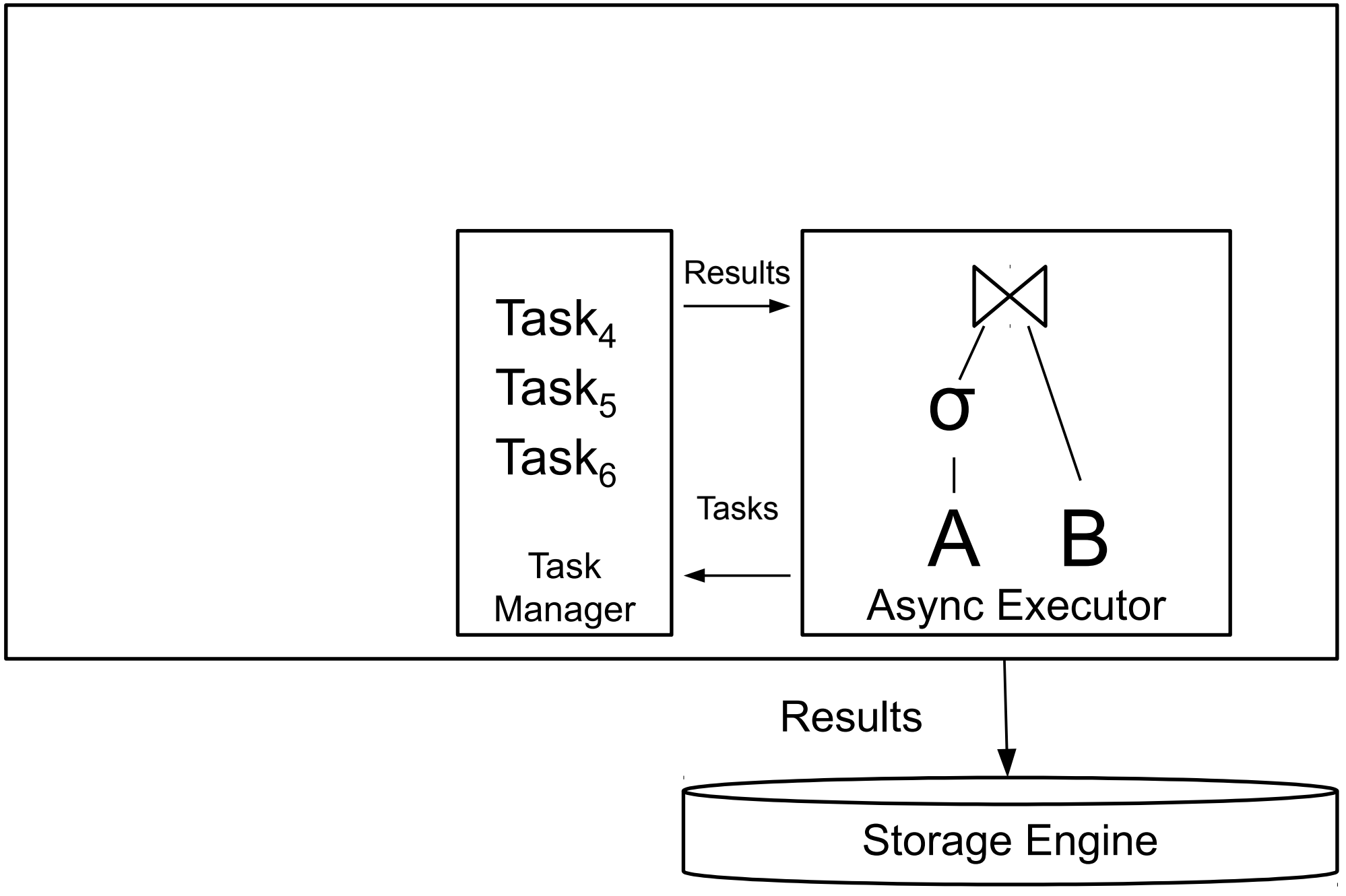


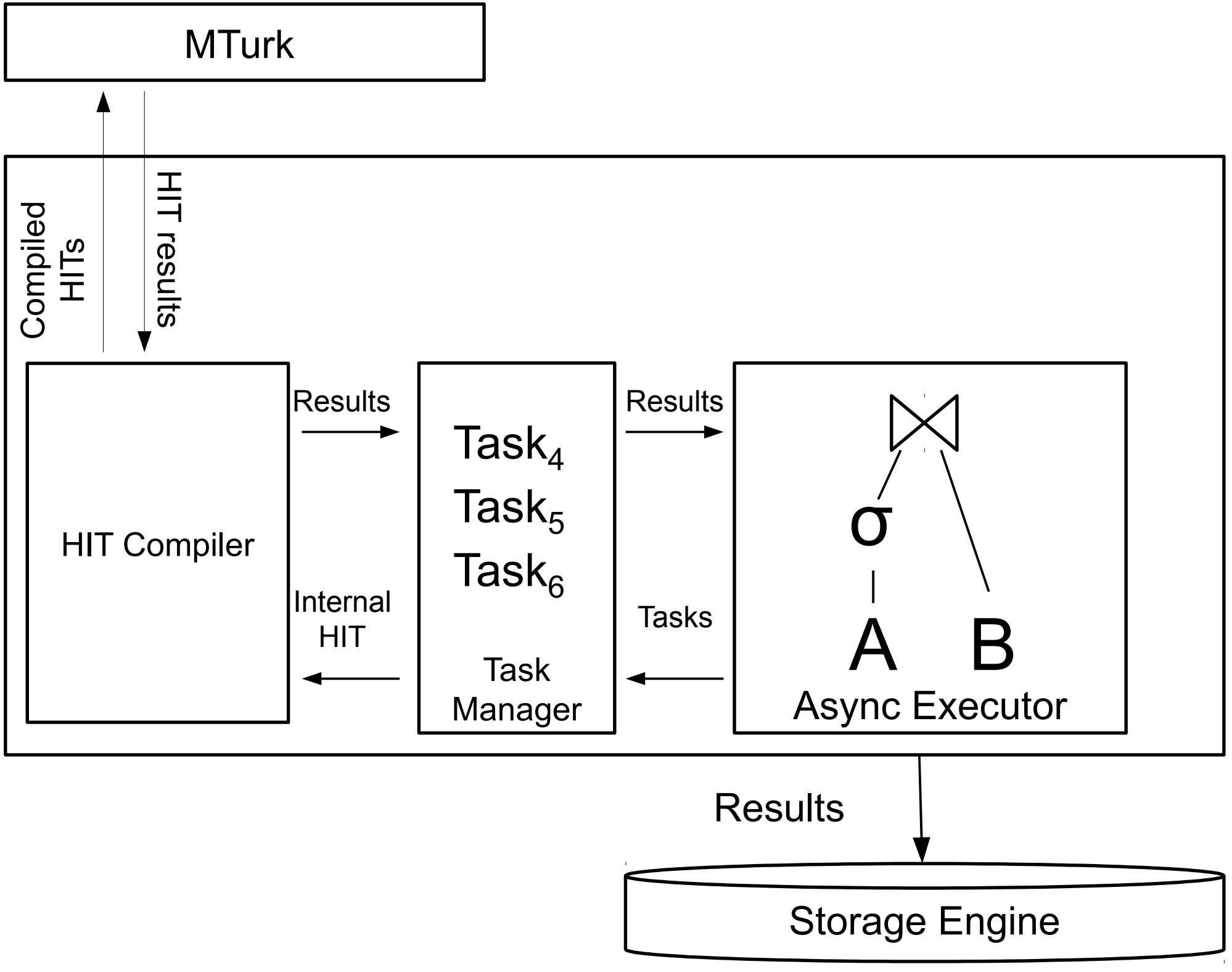
Haiti Join

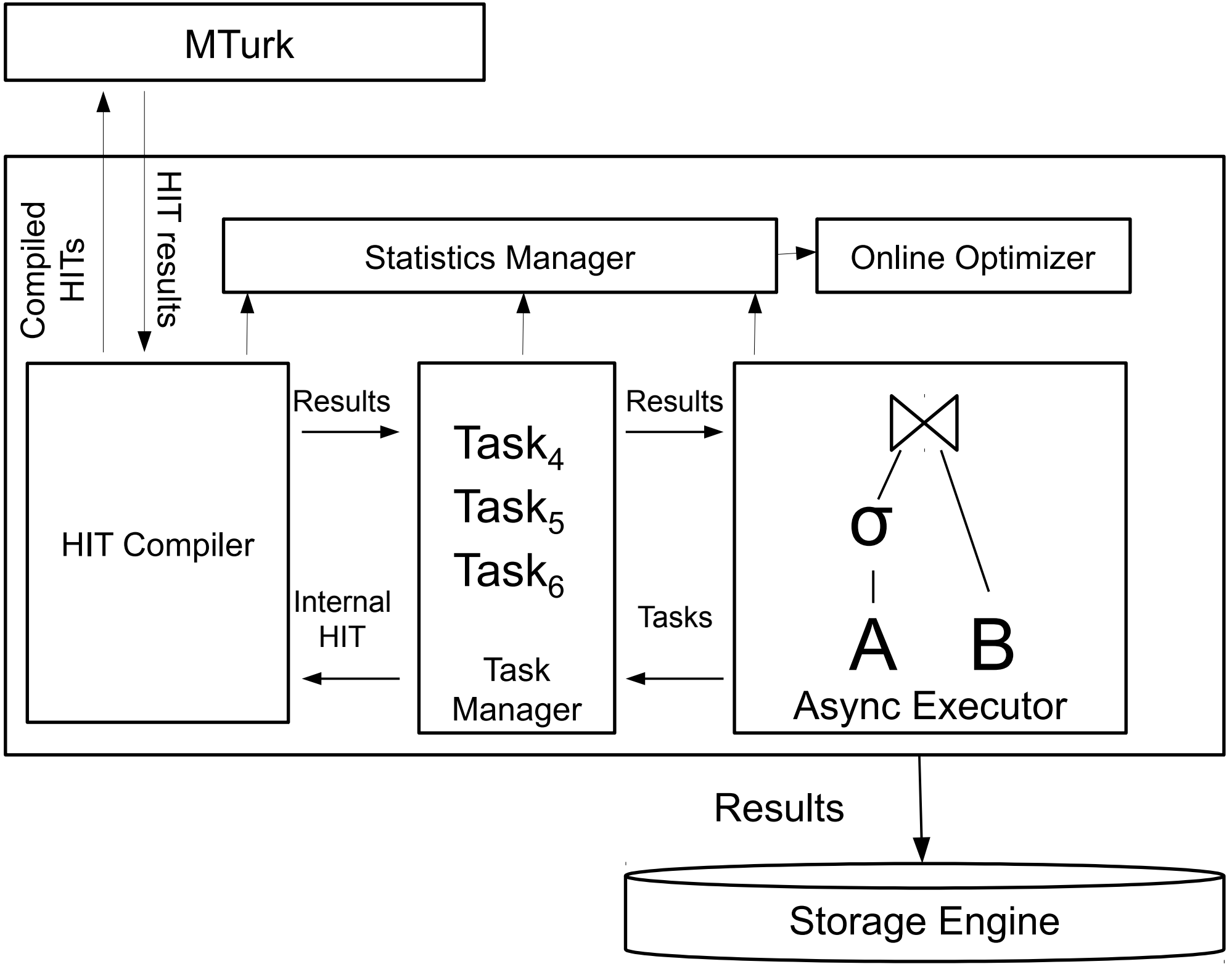
```
SELECT survivor.name
FROM survivor, missing
WHERE majorityVote(samePerson(survivor.img,
                              missing.img))
```

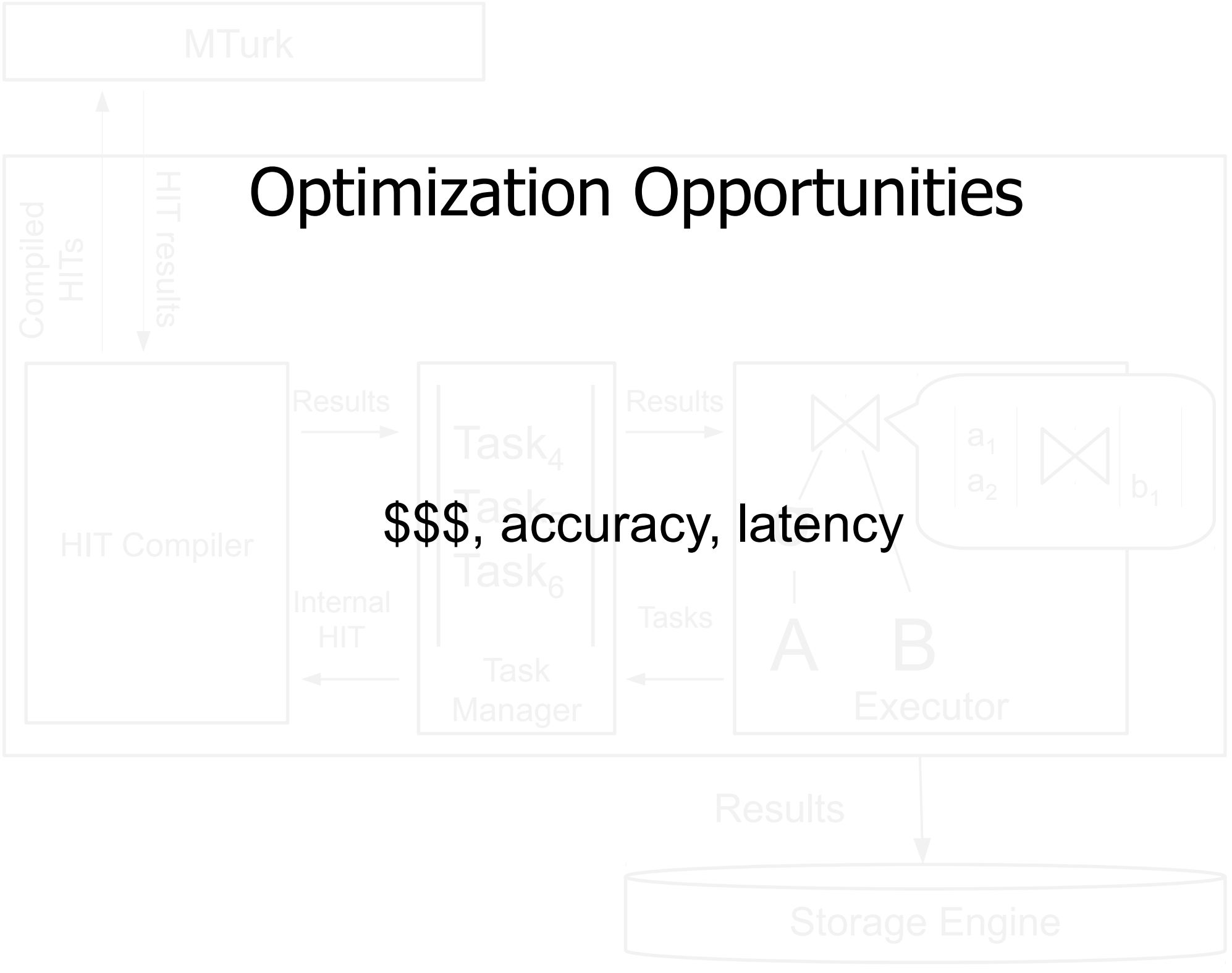
Qurk System Design




Task4
Task5
Task6
Task Manager
Results
Tasks
σ
A
B
Async Executor
Results
Storage Engine

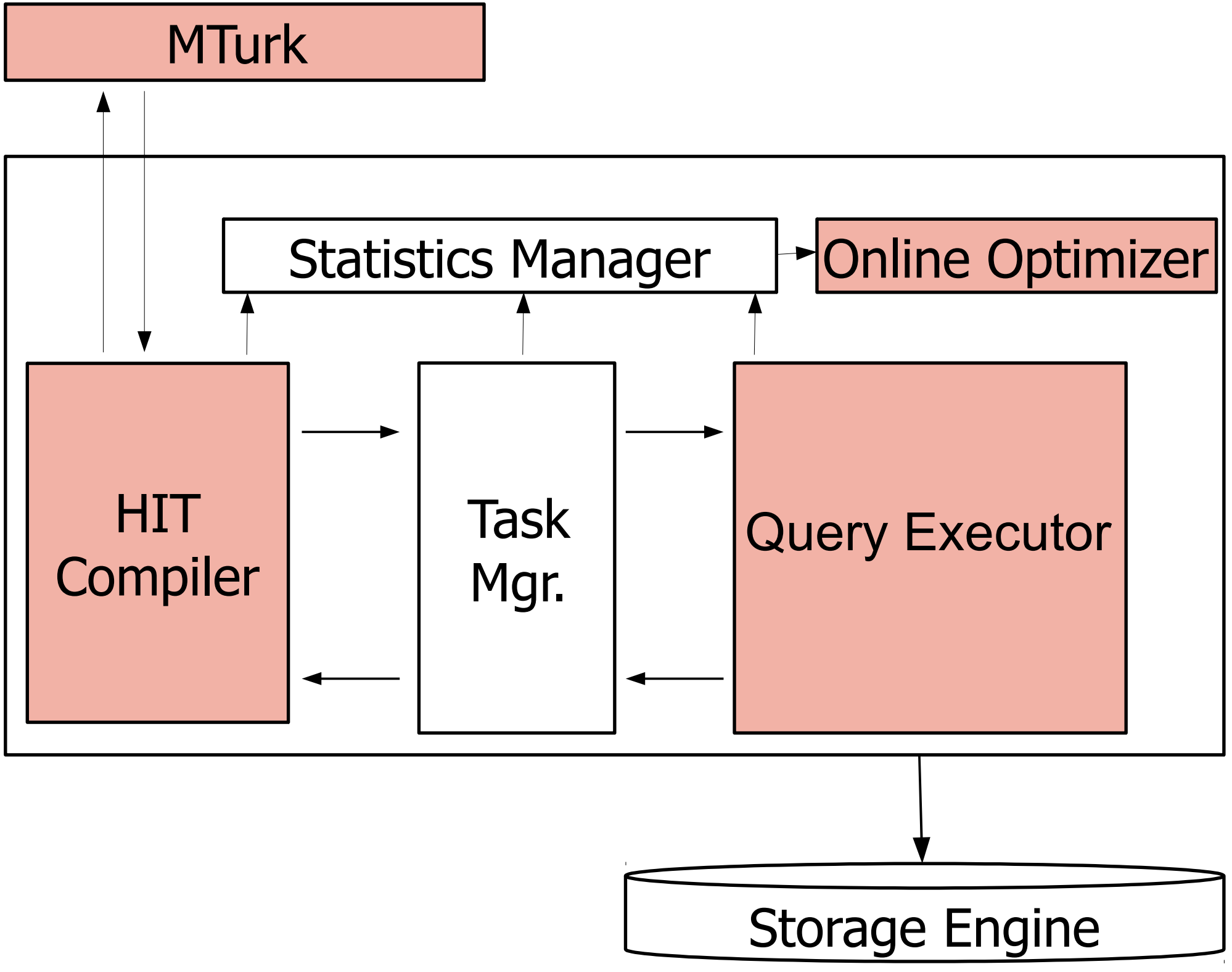

MTurk
Compiled HITs
HIT results
HIT Compiler
Results
Internal HIT
Task4
Task5
Task6
Task Manager
Results
Tasks
σ
A
B
Async Executor
Results
Storage Engine

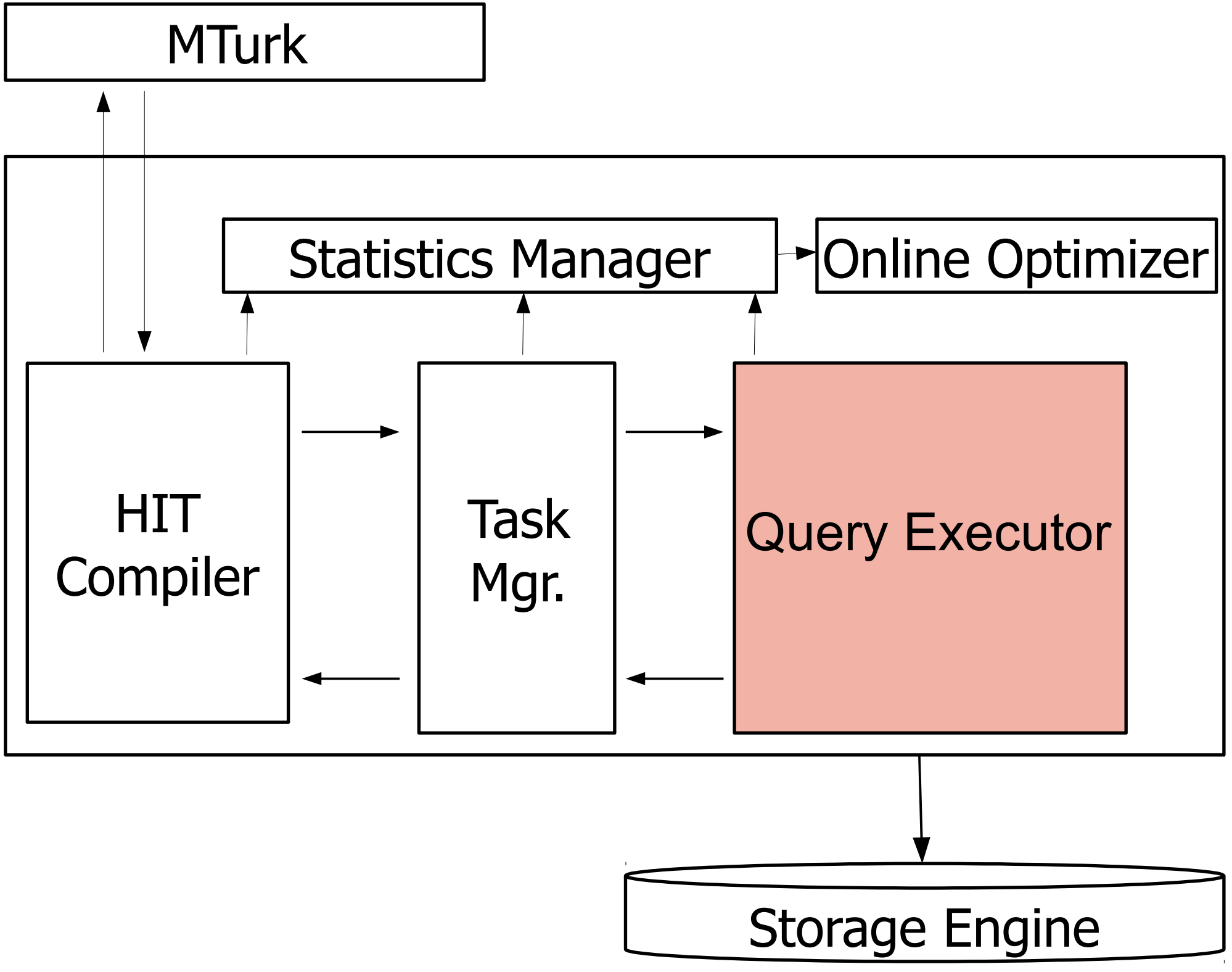

MTurk
Compiled HITs
HIT results
Statistics Manager
Online Optimizer
HIT Compiler
Results
Internal HIT
Task4
Task5
Task6
Task Manager
Results
Tasks
σ
A
B
Async Executor
Results
Storage Engine



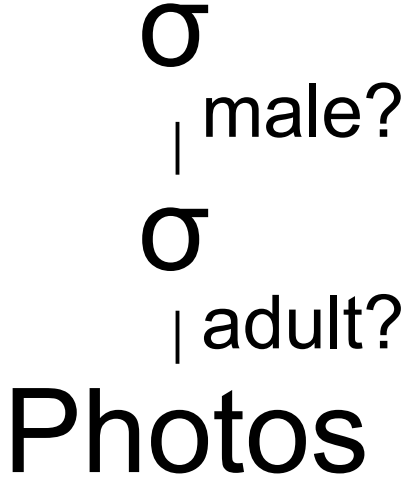
Optimization Opportunities

$$$, accuracy, latency


MTurk
Statistics Manager
Online Optimizer
HIT Compiler
Task Mgr.
Query Executor
Storage Engine


MTurk
Statistics Manager
Online Optimizer
HIT Compiler
Task Mgr.
Query Executor
Storage Engine

Combine Operators



Combine Operators

σ
| male?
σ
| adult?
Photos

Is this a picture of an adult? ○ Yes ○ No

Combine Operators

σ
| male?
σ
| adult?
Photos

Is this a picture of an adult? ○ Yes ○ No

Combine Operators

$\sigma_{\text{male?}}$

$\sigma_{\text{adult?}}$

Photos

Is this a picture of an adult? ○ Yes ○ No

Is this a picture of a male? ○ Yes ○ No

Batch Tuples

σ
| male?
σ
| adult?
Photos

Turker 1

Is this a picture of an adult? ○ Yes ○ No
Is this a picture of a male? ○ Yes ○ No

Turker 2

Is this a picture of an adult? ○ Yes ○ No
Is this a picture of a male? ○ Yes ○ No

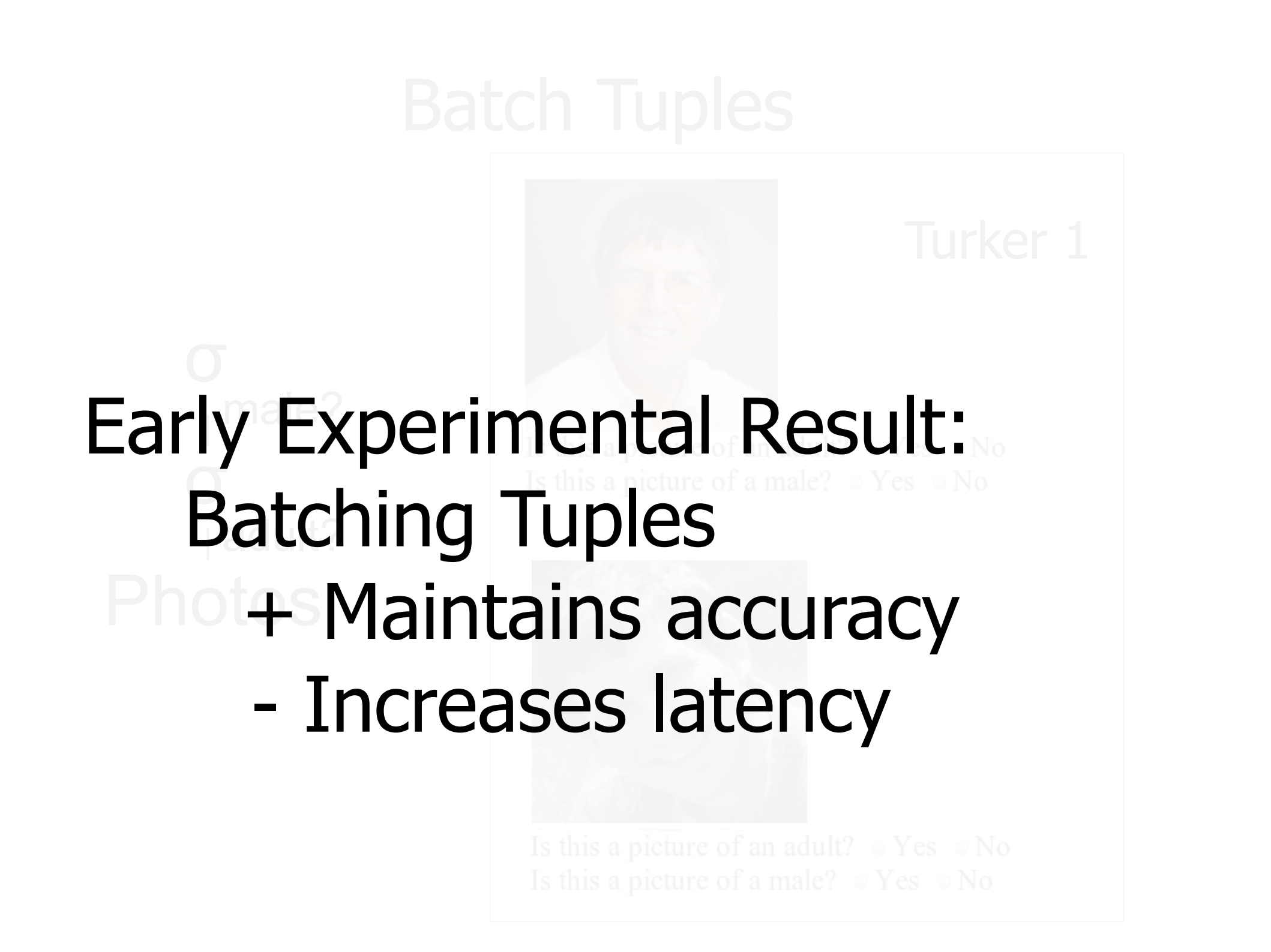
Batch Tuples

σ
| male?
σ
| adult?
Photos

Turker 1

Is this a picture of an adult? ○ Yes ○ No
Is this a picture of a male? ○ Yes ○ No

Is this a picture of an adult? ○ Yes ○ No
Is this a picture of a male? ○ Yes ○ No



Early Experimental Result: Batching Tuples

+ Maintains accuracy

- Increases latency

Join Optimizations

- Avoid cross product

Join Optimizations

- Avoid cross product
- Technique 1: batching

Join Optimizations

- Avoid cross product
- Technique 1: batching
- Technique 2: join heuristics reduce search space
  - e.g., image equi-join: gender, ethnicity match

Turker Fatigue

MWXMNMWNMNWMNXMN ○ Has V ○ No V

XWMXNWMNXXWMNWX ○ Has V ○ No V

NXNMXWNXWMXNWMN ○ Has V ○ No V

WMNWXNXWMWXMWXN ○ Has V ○ No V

NWXNXWMWXMWXMNW ○ Has V ○ No V

NMWNMNWMNXMNXNX ○ Has V ○ No V

XWMWXMWXMNMWNXW ○ Has V ○ No V

NXMNXNMXVWNXWM ○ Has V ○ No V

Turker Fatigue

MWXMNMWNMNWMNXMN ○ Has V ○ No V
XWMXNWMNXXWMNWX ○ Has V ○ No V
NXNMXWNXWMXNWMN ○ Has V ○ No V
WMNWXNXWMWXMWXN ○ Has V ○ No V
NWXNXWMWXMWXMNW ○ Has V ○ No V
NMWNMNWMNXMNXNX ○ Has V ○ No V
XWMWXMWXMNMWNXW ○ Has V ○ No V
NXMNXNMXVWNXWM ○ Has V ○ No V

Skew Correction

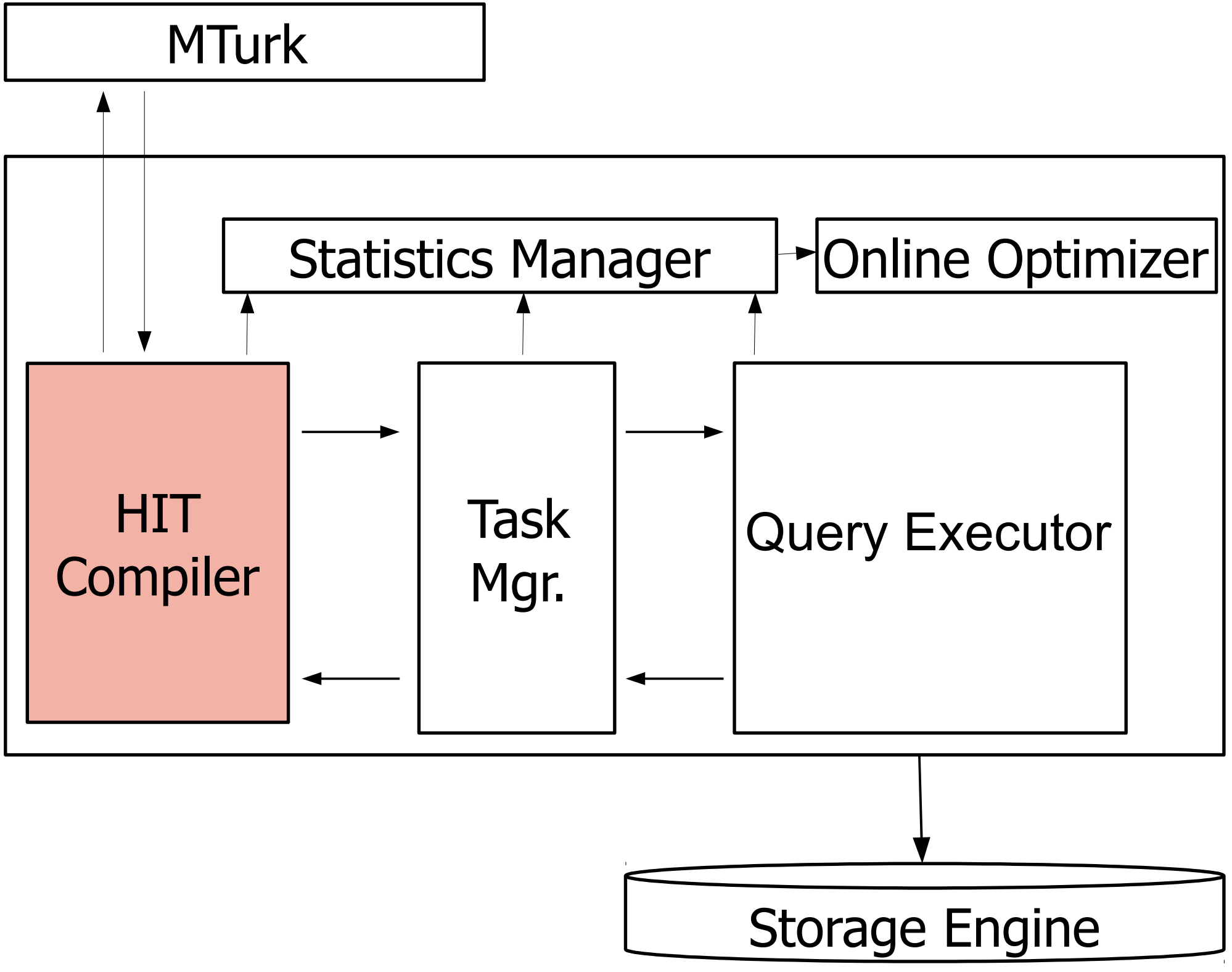
MWXMNMWNMNWMM ○ Has V ○ No V
XWMXNWMXVXXWMN ○ Has V ○ No V
NXNMXWNXWMXNW ○ Has V ○ No V
WMNWXVWXWMWXMW ○ Has V ○ No V
NWXNXWMWXMWXMN ○ Has V ○ No V
NMWNMNWVMNXMNX ○ Has V ○ No V
XWMWXMWXMNMWNM ○ Has V ○ No V
NXMNXNMXVWNXWM ○ Has V ○ No V

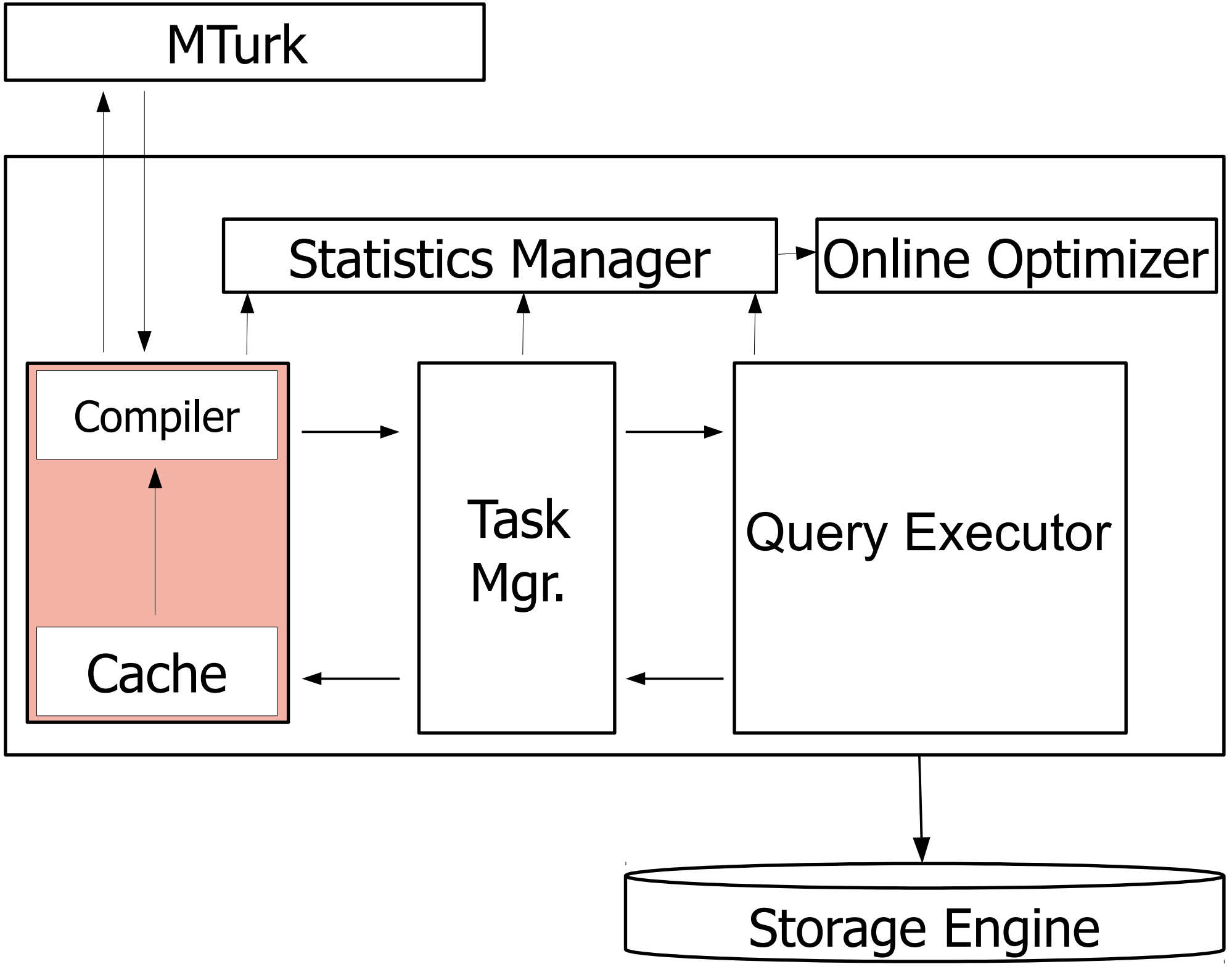
Early Experimental Result:

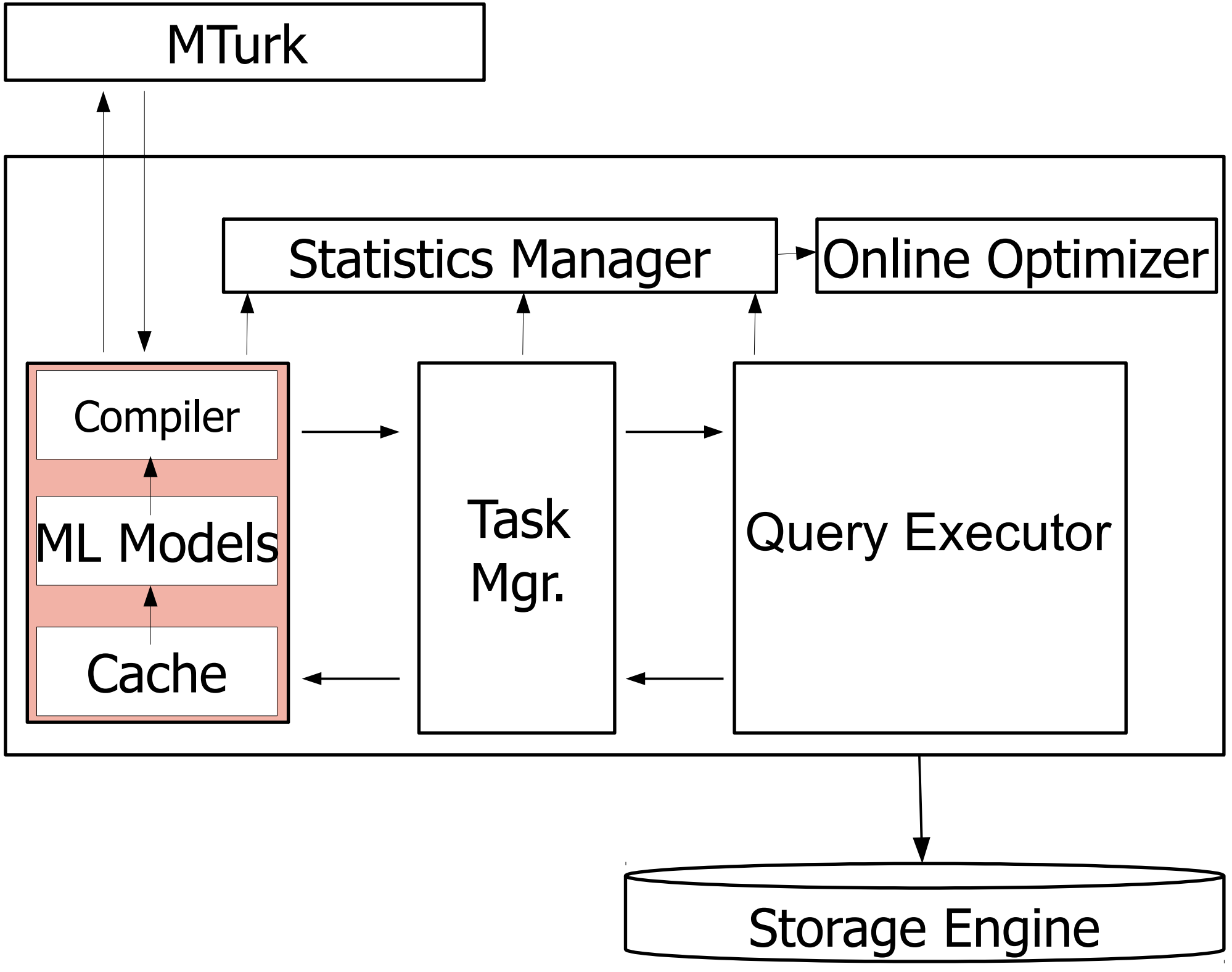
Skew Correction

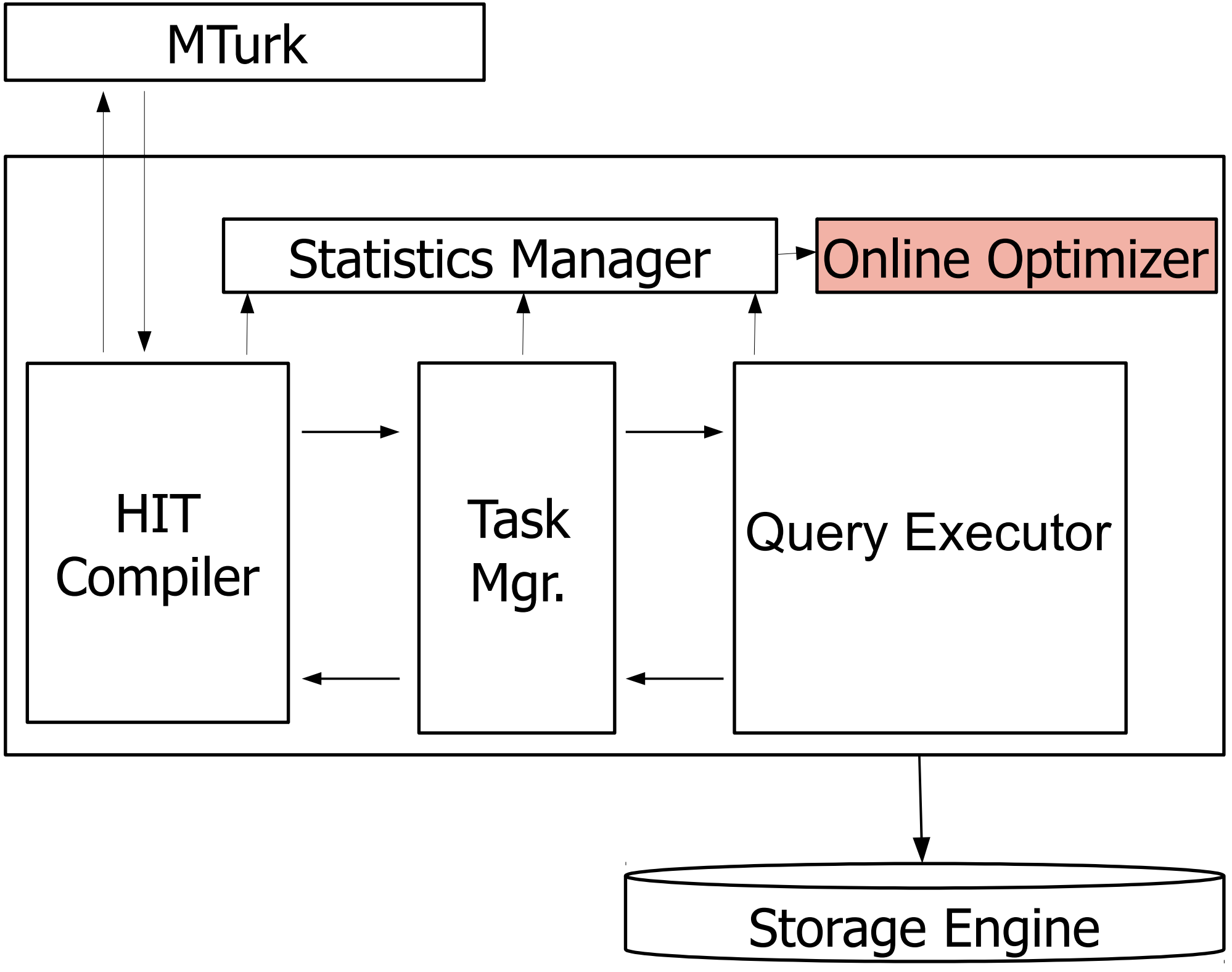
+ Improves Turker accuracy

- Increases number of HITs


MTurk
Statistics Manager
Online Optimizer
HIT Compiler
Task Mgr.
Query Executor
Storage Engine

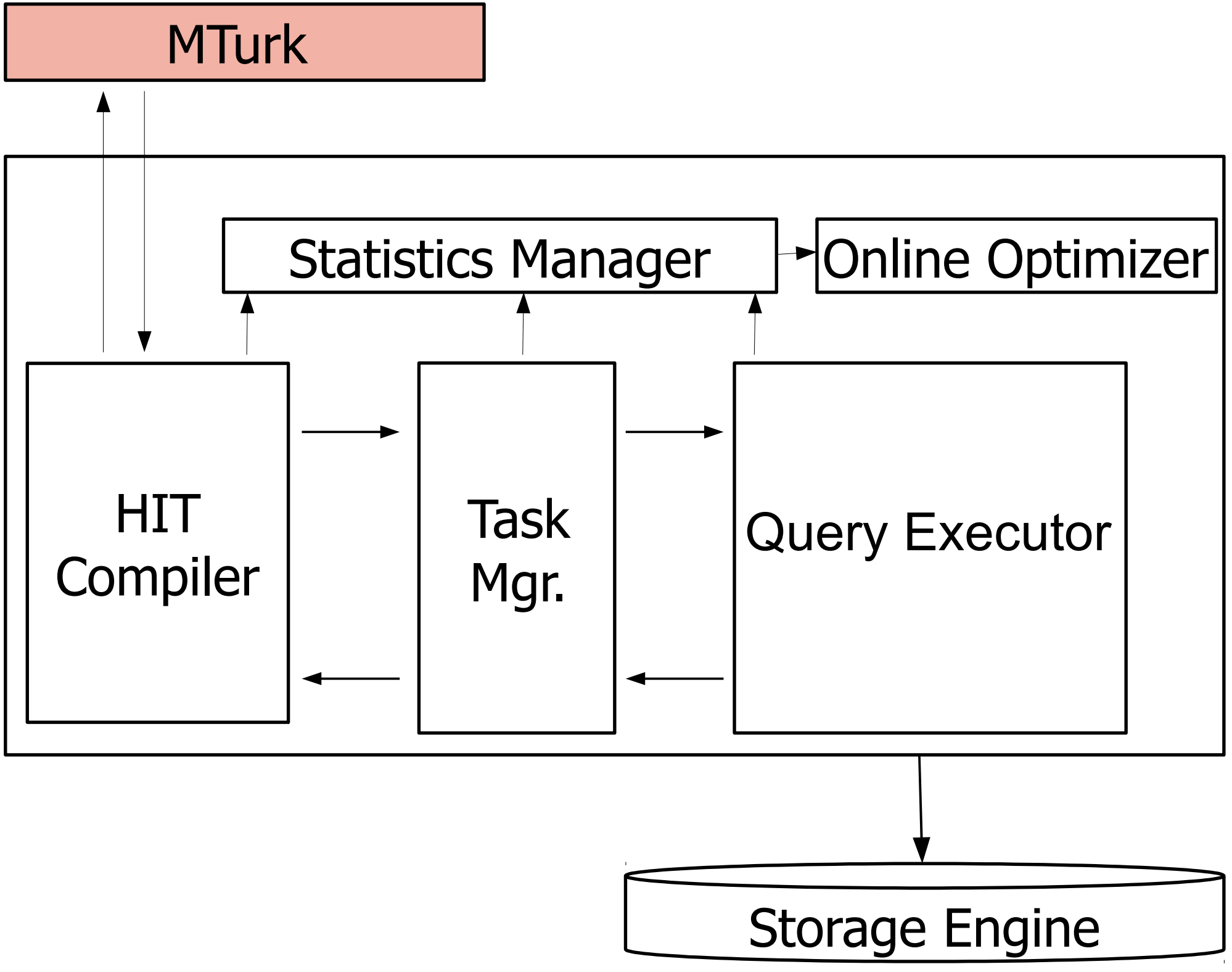

MTurk
Statistics Manager
Online Optimizer
Compiler
Cache
Task
Mgr.
Query Executor
Storage Engine


MTurk
Statistics Manager
Online Optimizer
Compiler
ML Models
Cache
Task Mgr.
Query Executor
Storage Engine


MTurk
Statistics Manager
Online Optimizer
HIT Compiler
Task Mgr.
Query Executor
Storage Engine

Online optimizations

- Batch size per HIT
- Price per HIT
  - Small effect on accuracy
  - Large effect on latency
- Turkers per HIT


MTurk
Statistics Manager
Online Optimizer
HIT Compiler
Task Mgr.
Query Executor
Storage Engine

Human Computation Platforms

- Paid Crowd (e.g., MTurk)
- Experts (e.g., Aardvark)
- Games (e.g., Image Labeling)

Qurk

- Human computation inside relational databases
- Data model: SQL + Lists
- Asynchronous system design
- Lots of optimization opportunities

Qurk

- Human computation inside relational databases
- Data model: SQL + Lists
- Asynchronous system design
- Lots of optimization opportunities
- Probably still can't con your way into CIDR

ask us for a demo!

marcua@csail.mit.edu/eugenewu@mit.edu

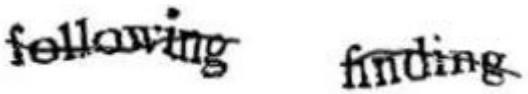
Image credits

- Clock: http://www.cs.wichita.edu/~vnambood/research.ht
- CAPTCHA: http://en.wikipedia.org/wiki/CAPTCHA

Statistics

- ~80k HITs available at any time
- ~$8K worth of work at any time
- ~1.2K projects at any time
- Source: http://www.mturk-tracker.com/general/

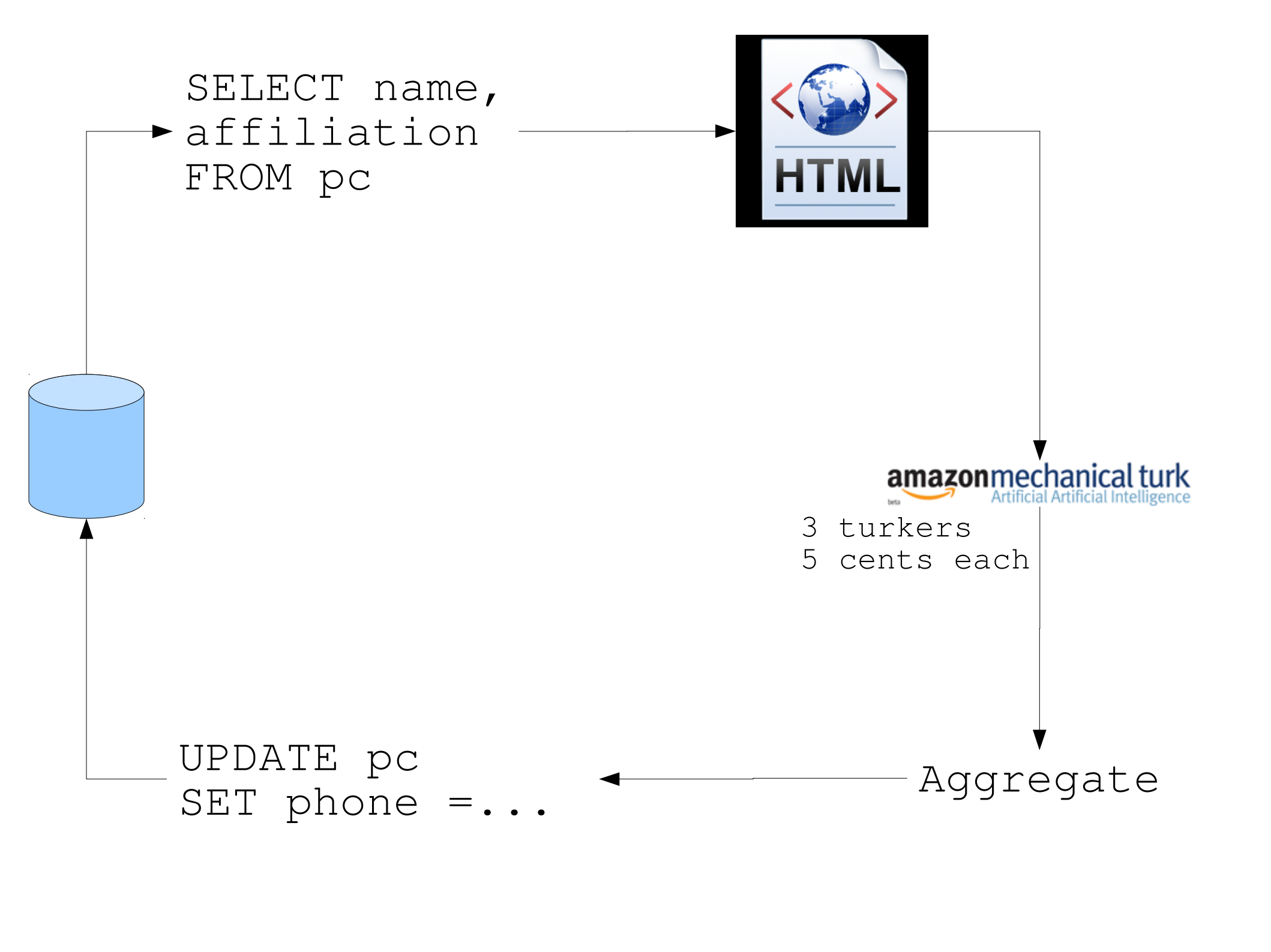
Other Human Computation Platforms

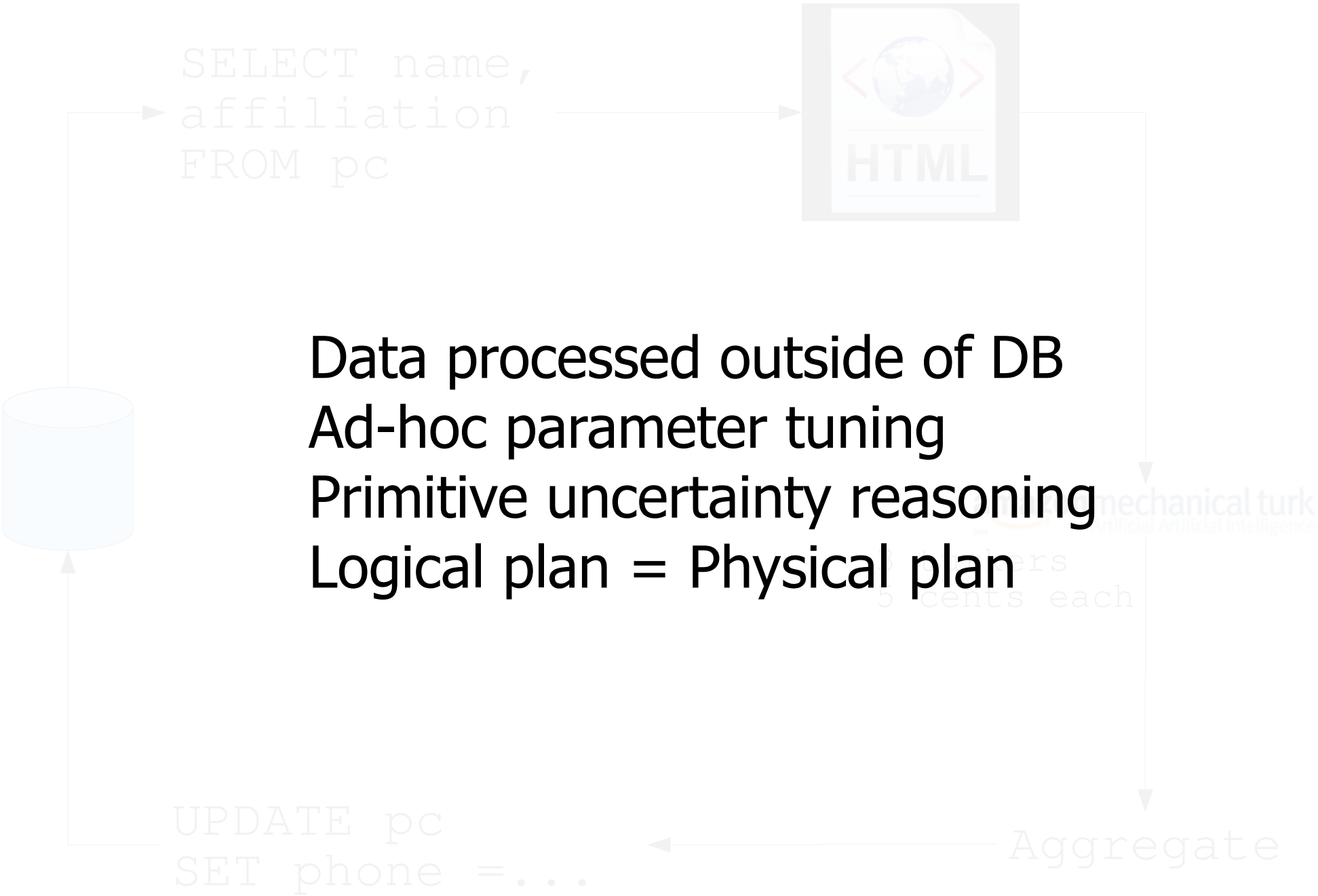
- Outsourced insurance claims processing
- Data journalism (Guardian)
- CAPTCHA ()
- Games with a purpose (image labeling)

Operator Implementations

- Non-blocking to allow pipelining (e.g., symmetric hash join)

- Join batching
- Rank by comparison, rating, or partial order


SELECT name,
affiliation
FROM pc
HTML
amazonmechanical turk
Artificial Artificial Intelligence
3 turkers
5 cents each
UPDATE pc
SET phone =...
Aggregate



Data processed outside of DB
Ad-hoc parameter tuning
Primitive uncertainty reasoning
Logical plan = Physical plan

ProPublica Reporting Network

ProPublica's Guide to Mechanical Turk

by Srinivas Rao and Amanda Michel
ProPublica, Oct. 15, 2010, 6:09 a.m.

Amazon Mechanical Turk – or mTurk – is an online marketplace, set up by the online shopping site Amazon, where anyone can hire workers to complete short, simple tasks

PRO PUBLICA Journalism in the public interest

About 3 years ago **Sean P. Aune** 9

Steve Fossett Search: Amazon Mechanical Turk Finds Leads

amazon mechanical turk beta Artificial Artificial Intelligence

As we reported previously, Amazon Mechanical Turk service is assisting in the search for adventurer Steve Fossett. Using quickly updated satellite imagery of the suspected crash area, people have been searching 256X256 squares for any sign of the aircraft.

There has now been an update on how the search is going, and it seems there have been some interesting leads. After each report is filed of a possible spotting, Mechanical Turk compares your report with any other reports for the same image. At that point, if the system determines there may be something of interest in the

Don't Miss: Dollars for Docs | Gulf Spill | Wall St. | TBI | Gas Drilling | Gitmo | Loan Mods | Stimul

PRO PUBLICA | Journalism in the public interest

Home | Our Investigations

ProPublica Reporti

ProPublica

by *Srinivas Rao* and *Amani*
ProPublica, Oct. 15, 2010, 6

Amazon Mechanical Tu
shopping site Amazon,

About 3 years ago *Sean P. Aune* 9

Steve Fossett Search: Amazon Mechanical Turk Finds Leads

Share

share

Do you like this story?

Like | Be the first of your frien

amazonmechanical tu
beta
Artificial Artificial Intellig

As we reported previously, Amazon Me
Fossett. Using quickly updated satellite
256X256 squares for any sign of the ai

There has now been an update on how
leads. After each report is filed of a pos
reports for the same image. At that poi

Tech | Gadgets | Mobile | Enterprise | GreenTech | CrunchBase | TechCrunch TV | The Crunchies | More

TechCrunch

It's not your dream
It's the one after th
Find

What's Hot: Android | Apple | Facebook | Google | Groupon | Microsoft | Zynga | Classics | CES 201

SpeakerText Automates And Crowdsources Video Transcripts (100 Beta Invites)

Erick Schonfeld
Sep 7, 2010

Like 49 Tweet 4 Digg

View Comments

SpeakerText BETA

One of the big problems with video on the Web is that other than the title, description and some meta tags, it is mostly invisible to Google and other search engines. One way to make video more SEO-friendly is to add transcriptions, but that can get expensive. An angel-funded startup called **SpeakerText** is (re)launching today with a very clever way to automate the transcription process and attach the full transcript as part of the video player in a drop down window. You can see an example of how this works below. And if you publish a lot of videos and want to try it out yourself, we have 100 **beta invites** (use the code: techcrunch).

Once a video is transcribed, it appears in a collapsible window below each player. Not only is

Ditch images, make it a list

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't relevant to a specific task. | Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't relevant to a specific task. | Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, as structure differences aren't important to the editing task. | Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. | Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Soylent's Shortn

Accenture Insurance BPO Services

Overview | Why Accenture | Specific Services

Overview

Leading insurers are increasingly turning to outsourcing to help them achieve high performance through operational efficiency.

Many insurers have already achieved notable success by outsourcing to Accenture Insurance Services, a specialized business process outsourcer with a unique combination of insurance and process optimization experience. Accenture Insurance Services helps clients simplify, automate and optimize their business processes to dramatically improve financial and operational performance.

By using mature platforms across multiple insurance processing units, we can deliver not only industry best practices, but also economies of scale. In addition, the Accenture Life Insurance Platform (formerly NaviSys) offers clients a full range of top-rated life policy processing solutions.

guardian.co.uk

Investigate your MP's expenses

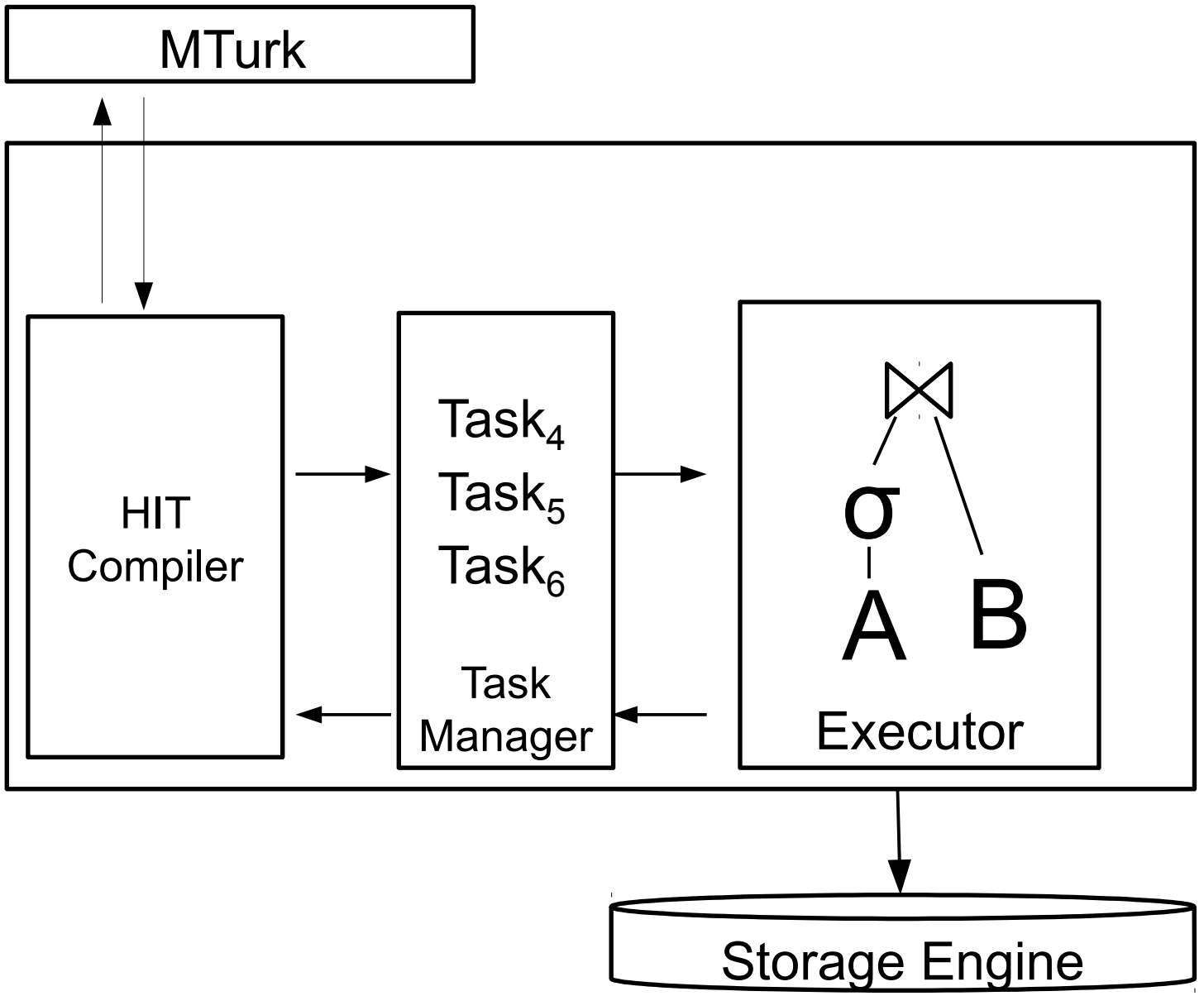
Join us in digging through the documents of MPs' expenses to identify individual claims, or documents that you think merit further investigation. You can work through your own MP's expenses, or just hit the button below to start reviewing. (Update, Fri pm: we now have a virtually complete set of expenses documents so you should be able to find your MP's) Already created an account? Log in here.

We have **458,832** pages of documents. **27,486** of you have reviewed **222,875** of them. Only **235,957** to go...

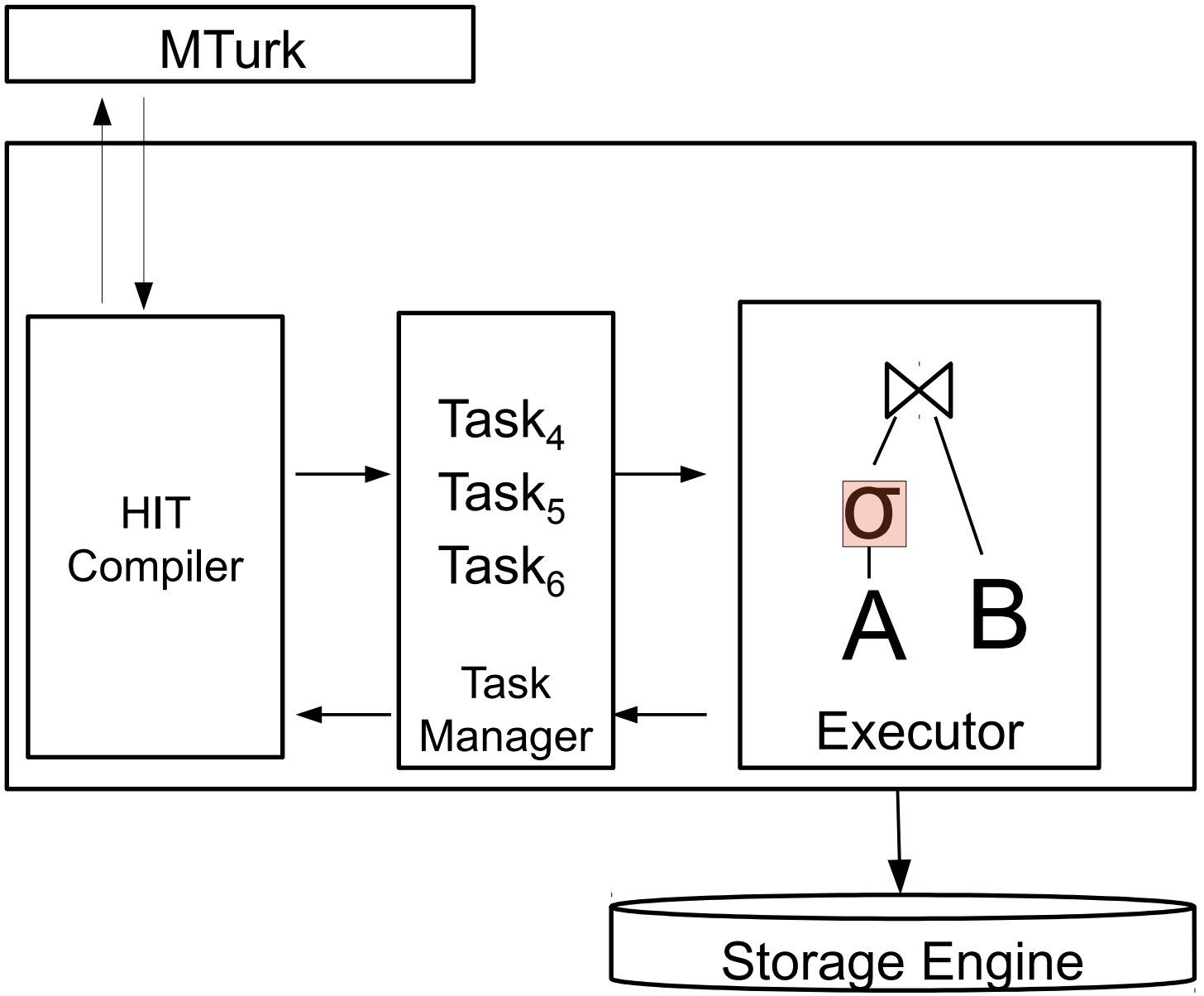
Start reviewing

Wrap this into mechanical turk slide, do as text

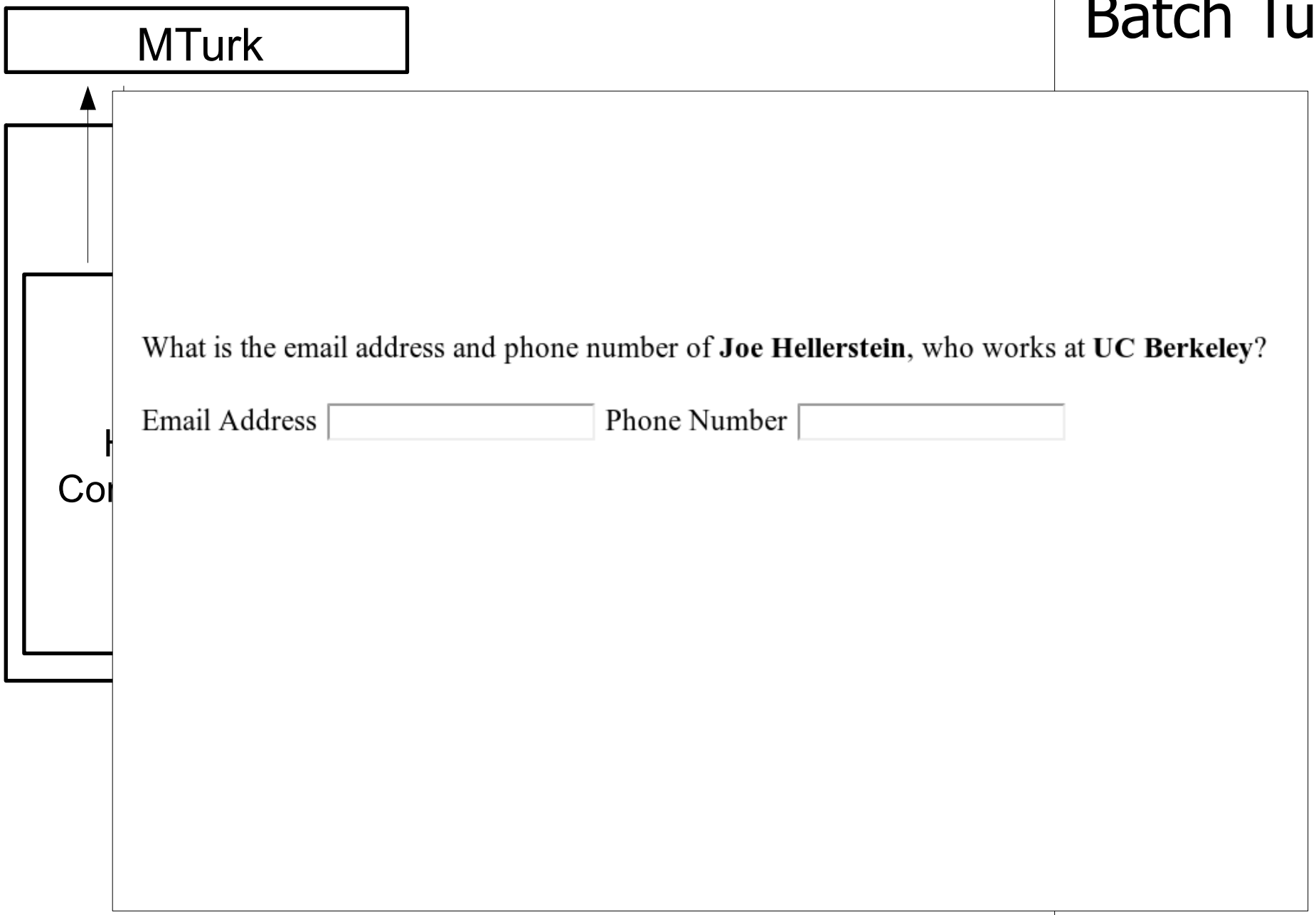
(want more about information integration, data cleaning)


MTurk
HIT
Compiler
Task4
Task5
Task6
Task
Manager
σ
A
B
Executor
Storage Engine

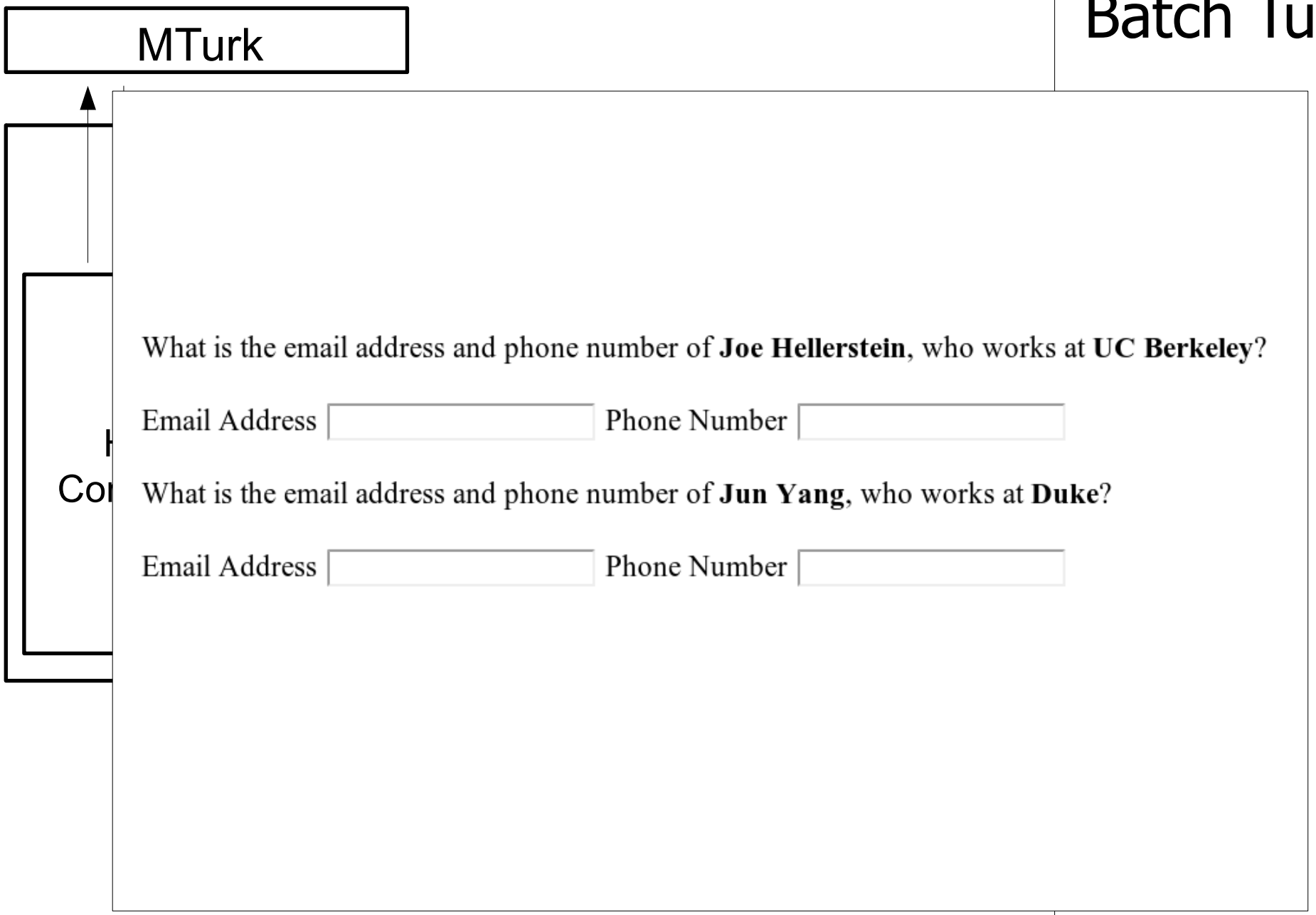
Batch Tuples



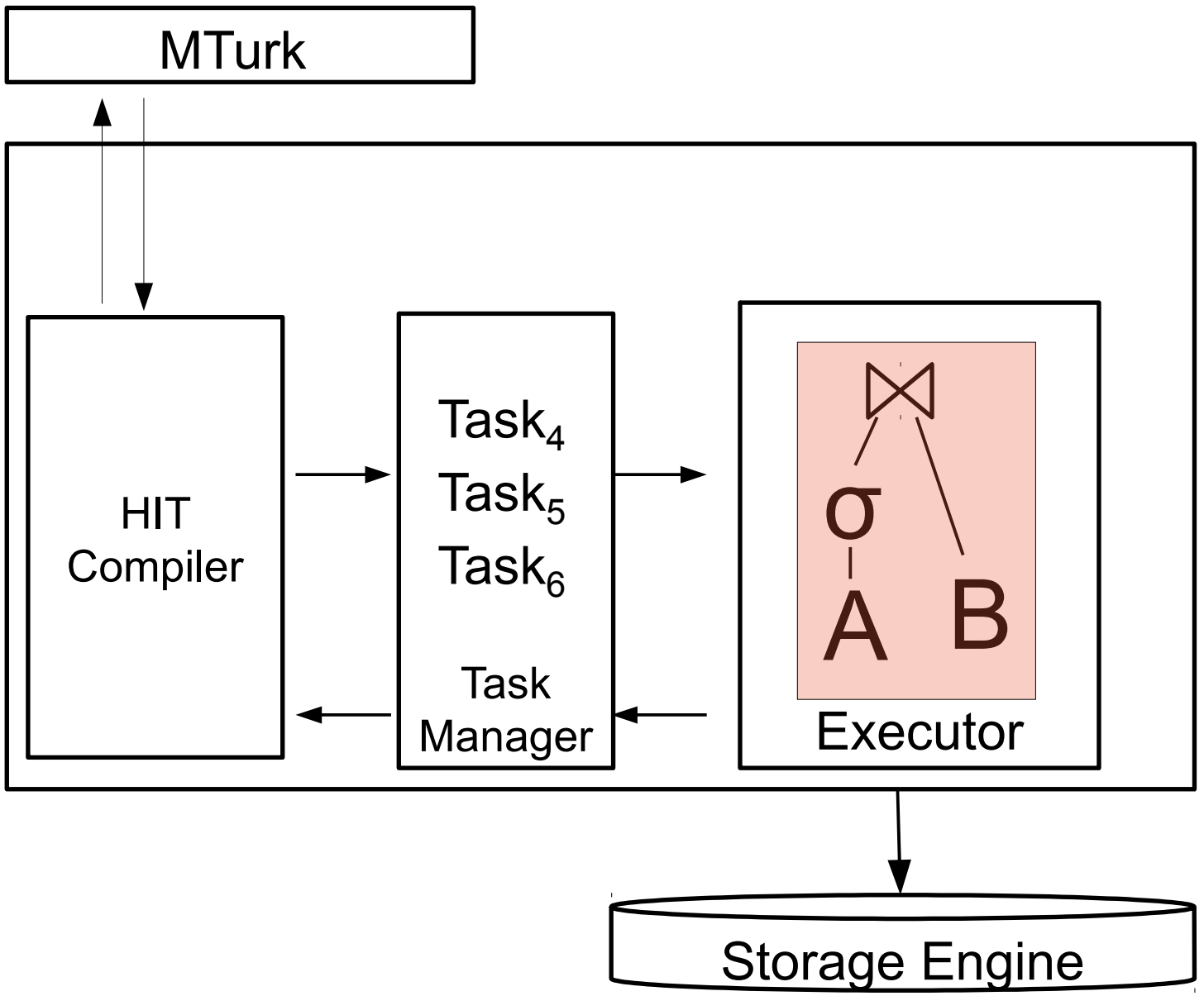
Batch Tuples



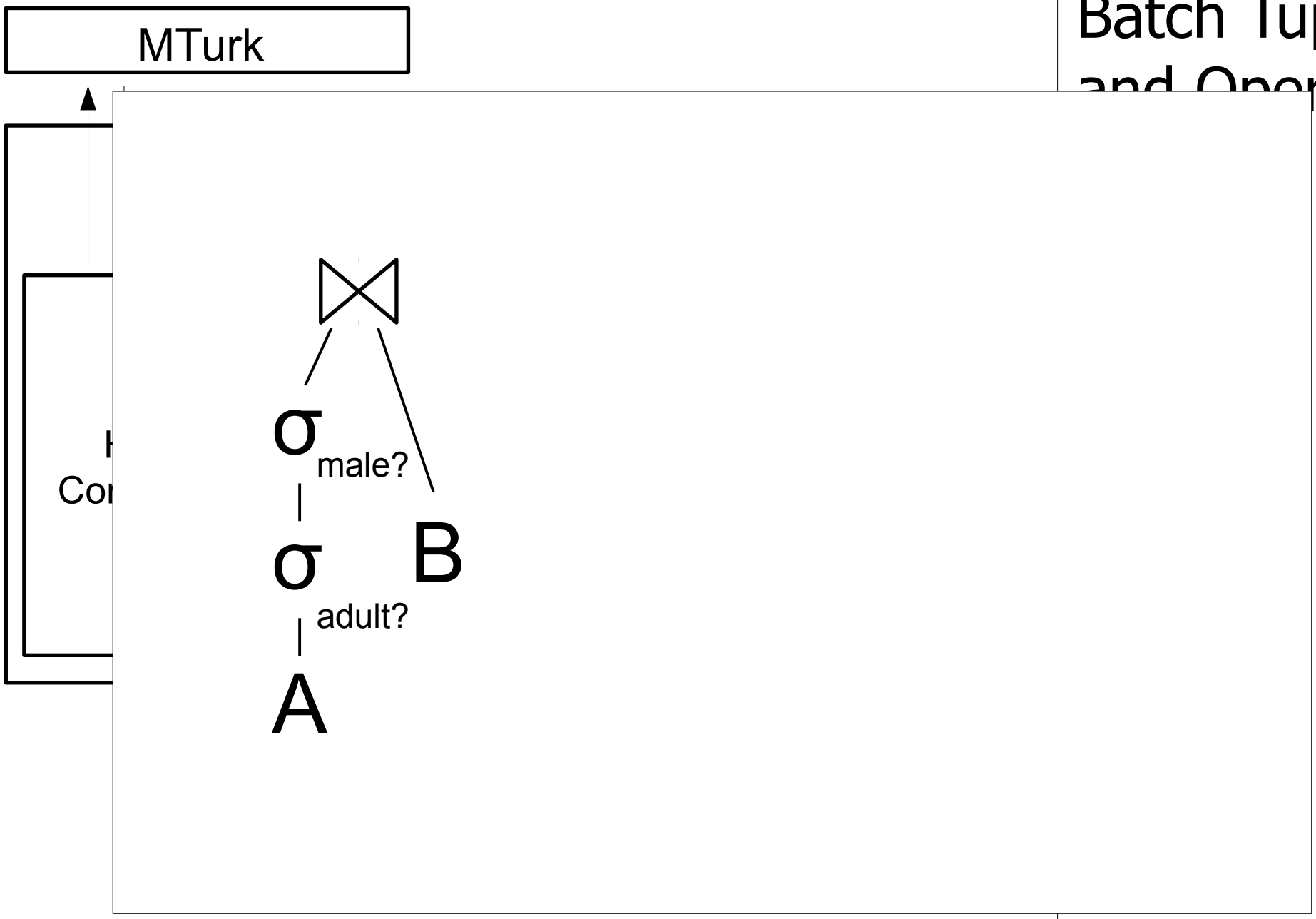
Batch Tuples



Batch Tuples and Operators

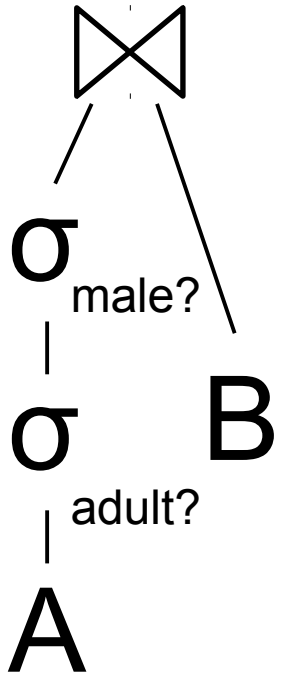


Batch Tuples and Operators



Batch Tuples and Operators

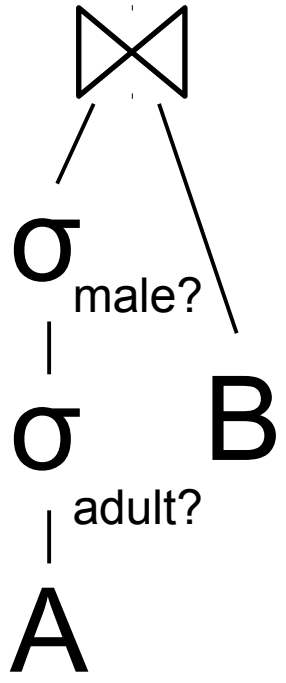
MTurk



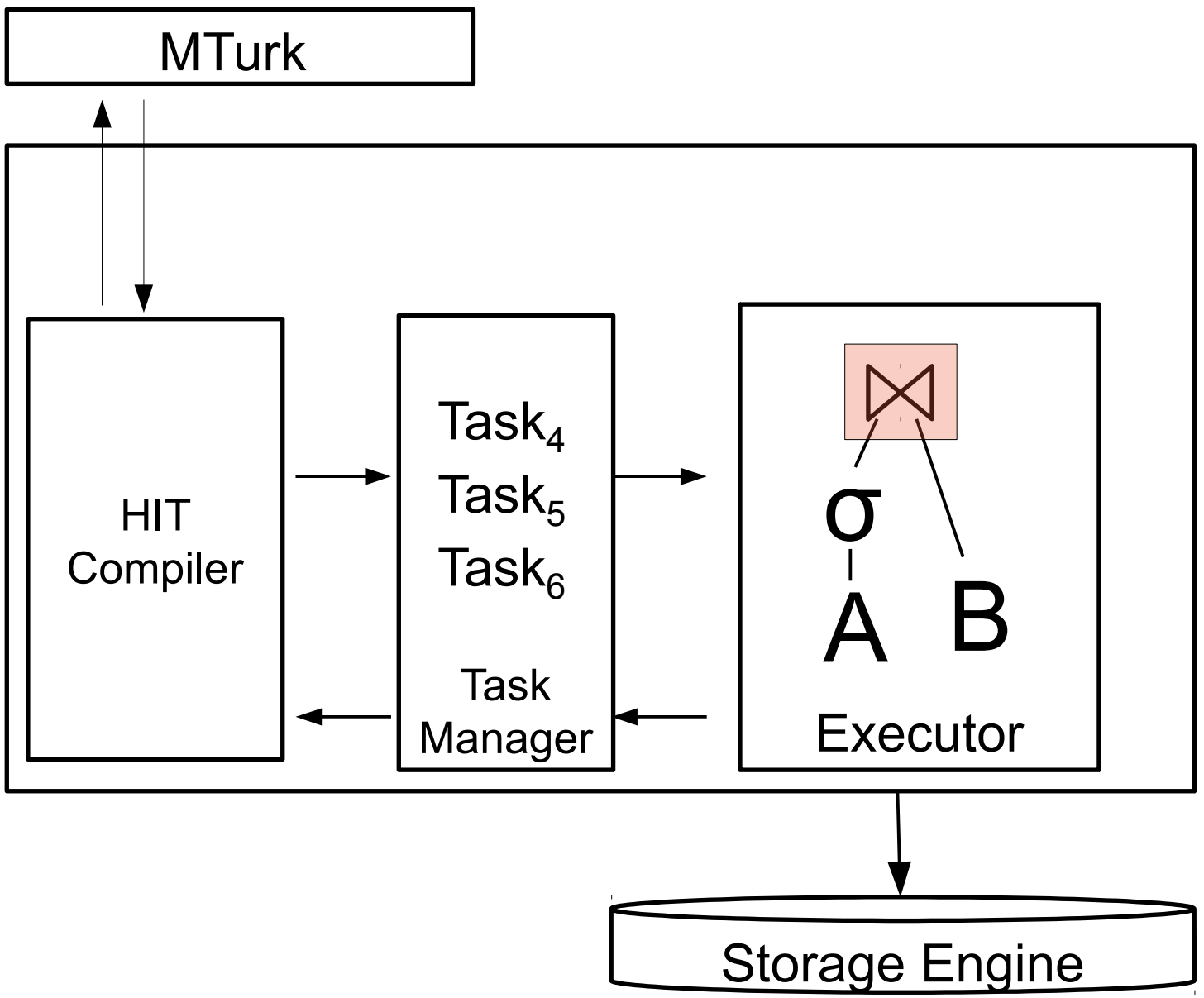
Is this a picture of an adult? Yes No

Batch Tuples and Operators

MTurk

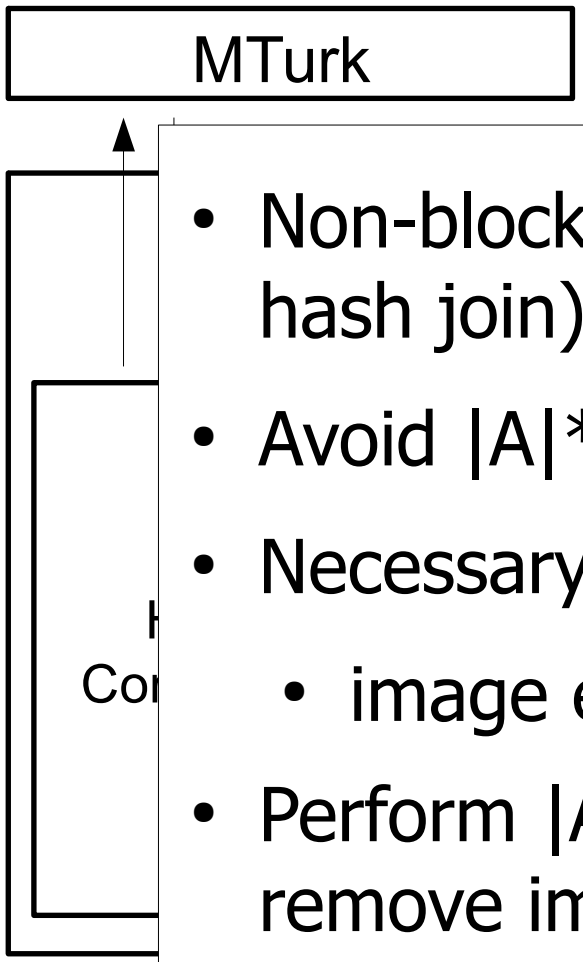


Is this a picture of an adult? Yes No
Is this a picture of a male? Yes No



Batch Tuples and Operators

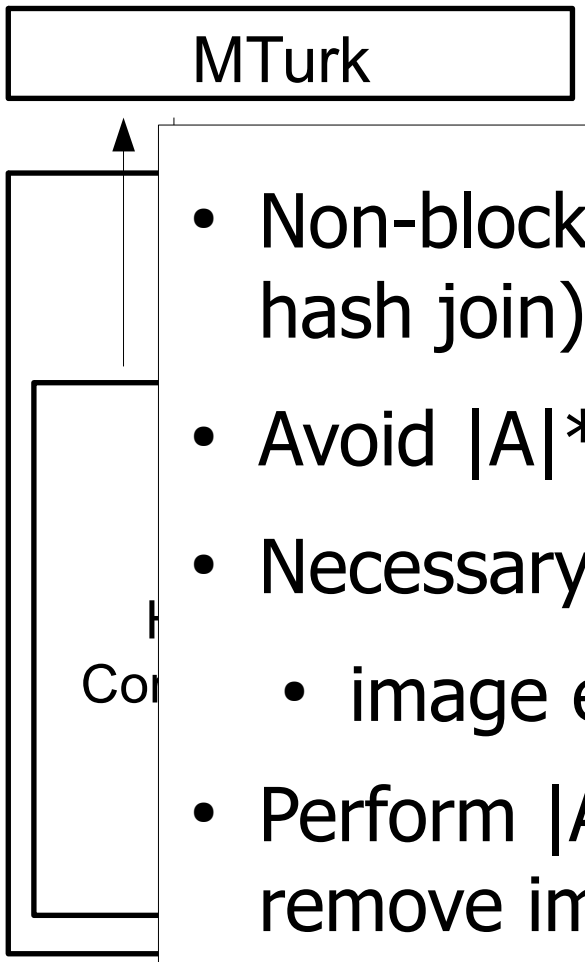
Operator Optimizations



Batch Tuples and Operators

tions

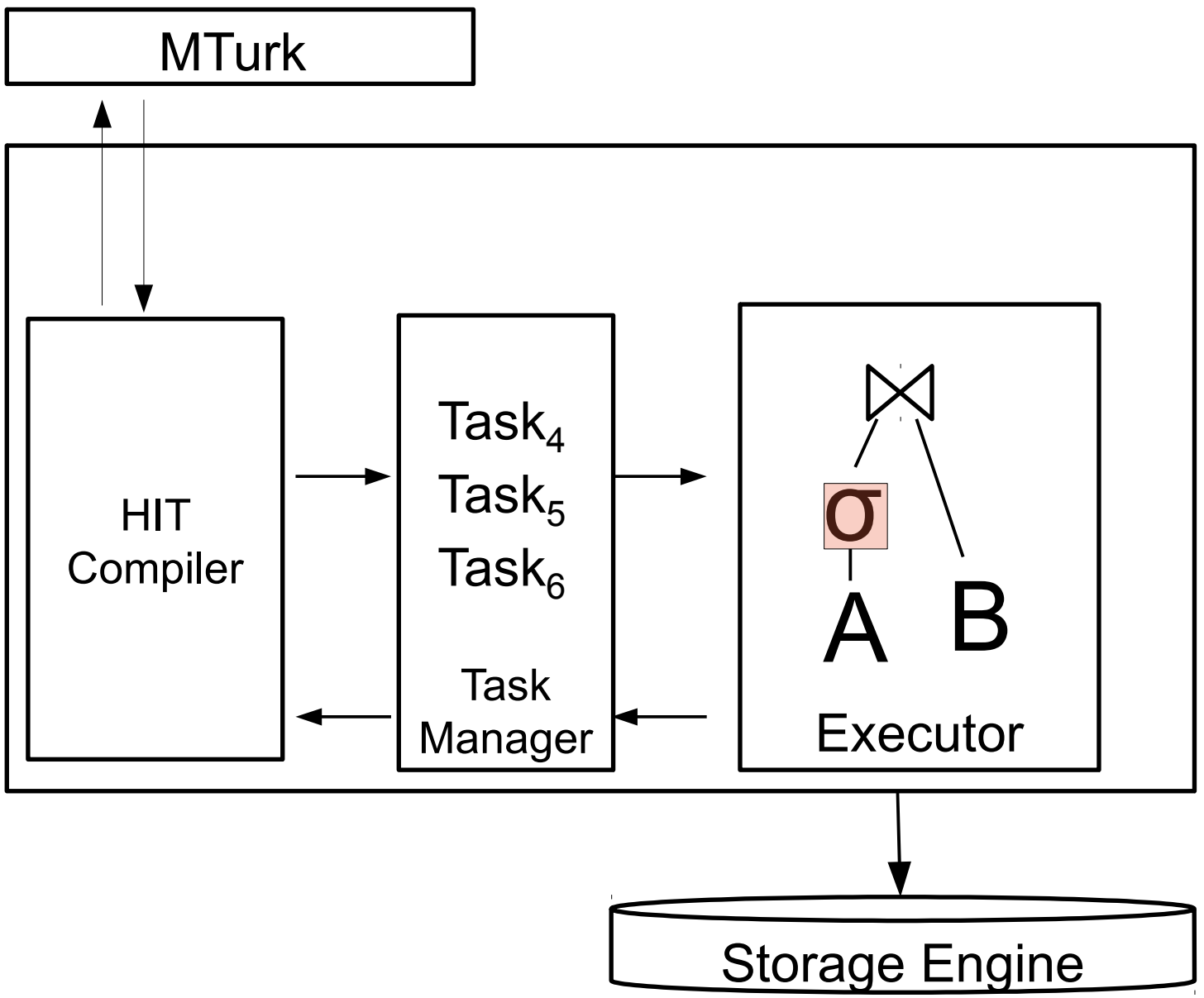
- Non-blocking for pipelining (e.g. symmetric hash join)
- Avoid |A|*|B| join cost
- Necessary conditions for equality
  - image equi-join: gender, ethnicity match
- Perform |A| + |B| scans to find predicates, remove impossible join pairs



Batch Tuples and Operators

tions

- Non-blocking for pipelining (e.g. symmetric hash join)
- Avoid |A|*|B| join cost
- Necessary conditions for equality
  - image equi-join: gender, ethnicity match
- Perform |A| + |B| scans to find predicates, remove impossible join pairs
- Alternative: cluster images



Batch Tuples and Operators

Operator Optimizations

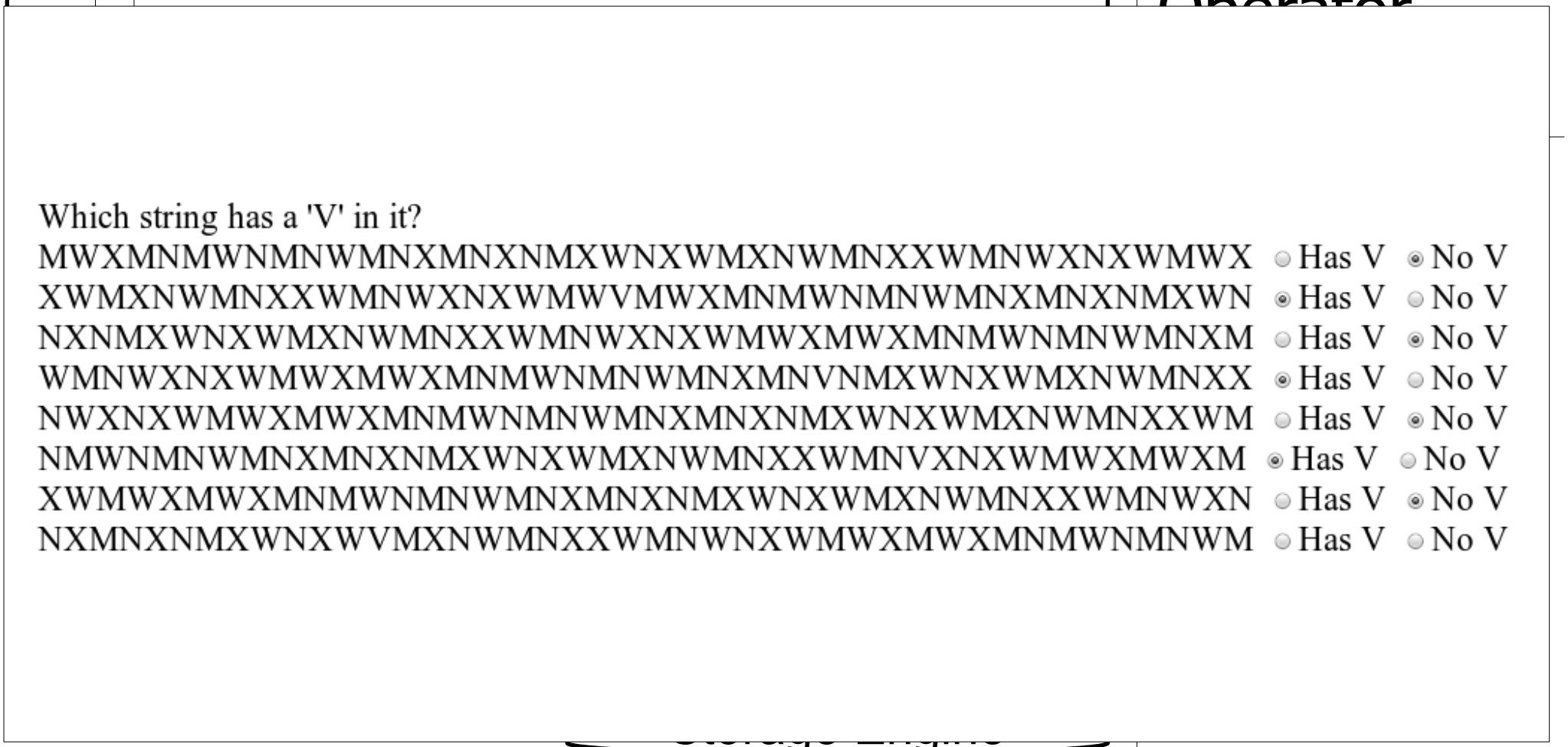
Selectivity Injection

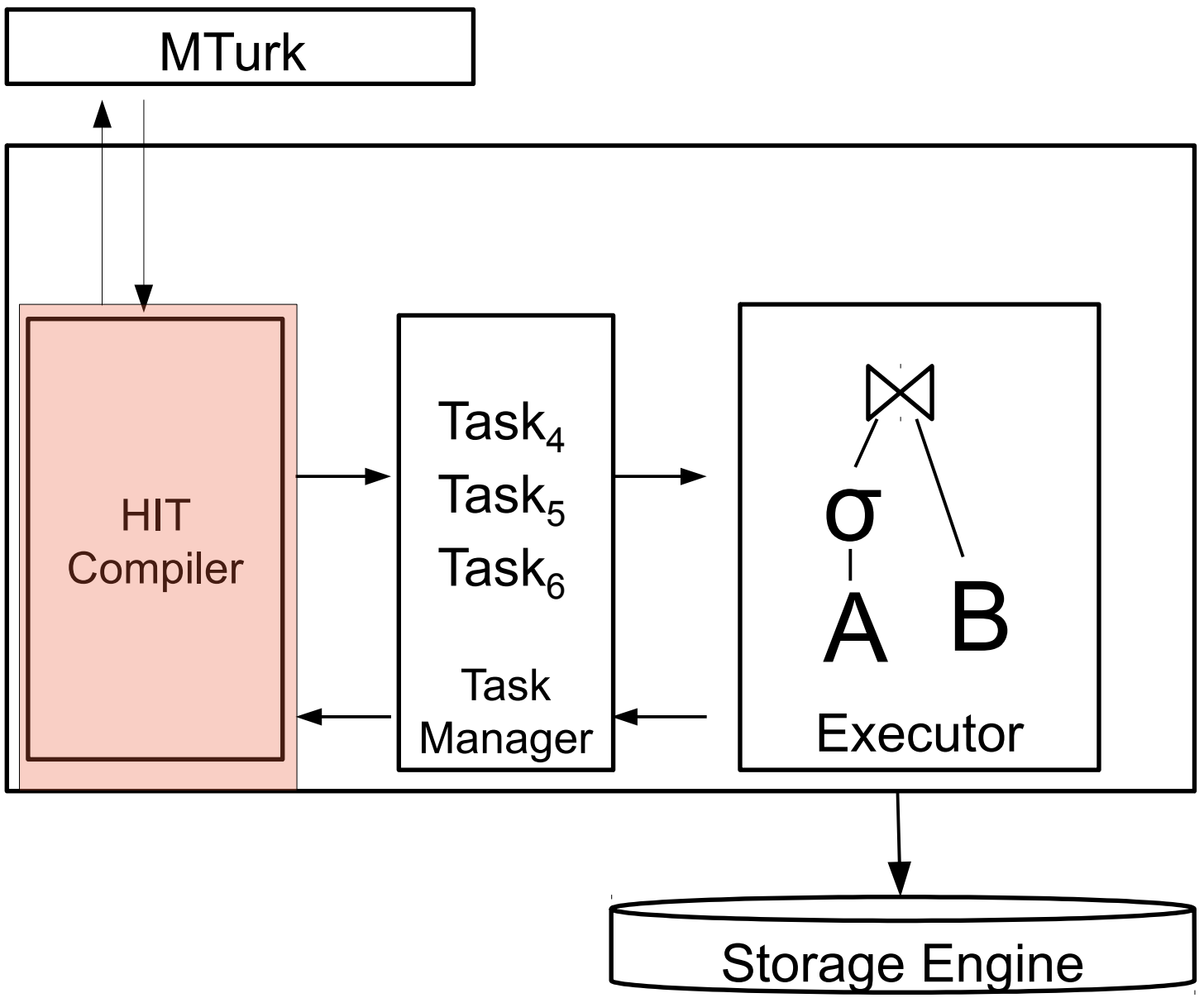

MTurk
Batch Tuples and Operators
Operator
Which string has a 'V' in it?
MWXMNMWNMNWMNXMNXNMXWNXWMXNWMNXXWMNWXNXWMWX Has V No V
XWMXNWMNXXWMNWXNXWMWMMWXMNMWNMNWMNXMNXNMXWN Has V No V
NXNMXWNXWMXNWMNXXWMNWXNXWMWXMWXMNMWNMNWMNXM Has V No V
WMNWXNXWMWXMWXMNMWNMNWMNXMNMNMXWNXWMXNWMNXX Has V No V
NWXNXWMWXMWXMNMWNMNWMNXMNXNMXWNXWMXNWMNXXWM Has V No V
NMWNMNWMNXMNXNMXWNXWMXNWMNXXWMNMXNXWMWXMWXM Has V No V
XWMWXMWXMNMWNMNWMNXMNXNMXWNXWMXNWMNXXWMNWXN Has V No V
NXMNXNMXWNXWVMXNWMNXXWMNWNXWMWXMWXMNMWNMNWM Has V No V

MTurk

Batch Tuples and Operators

Operator



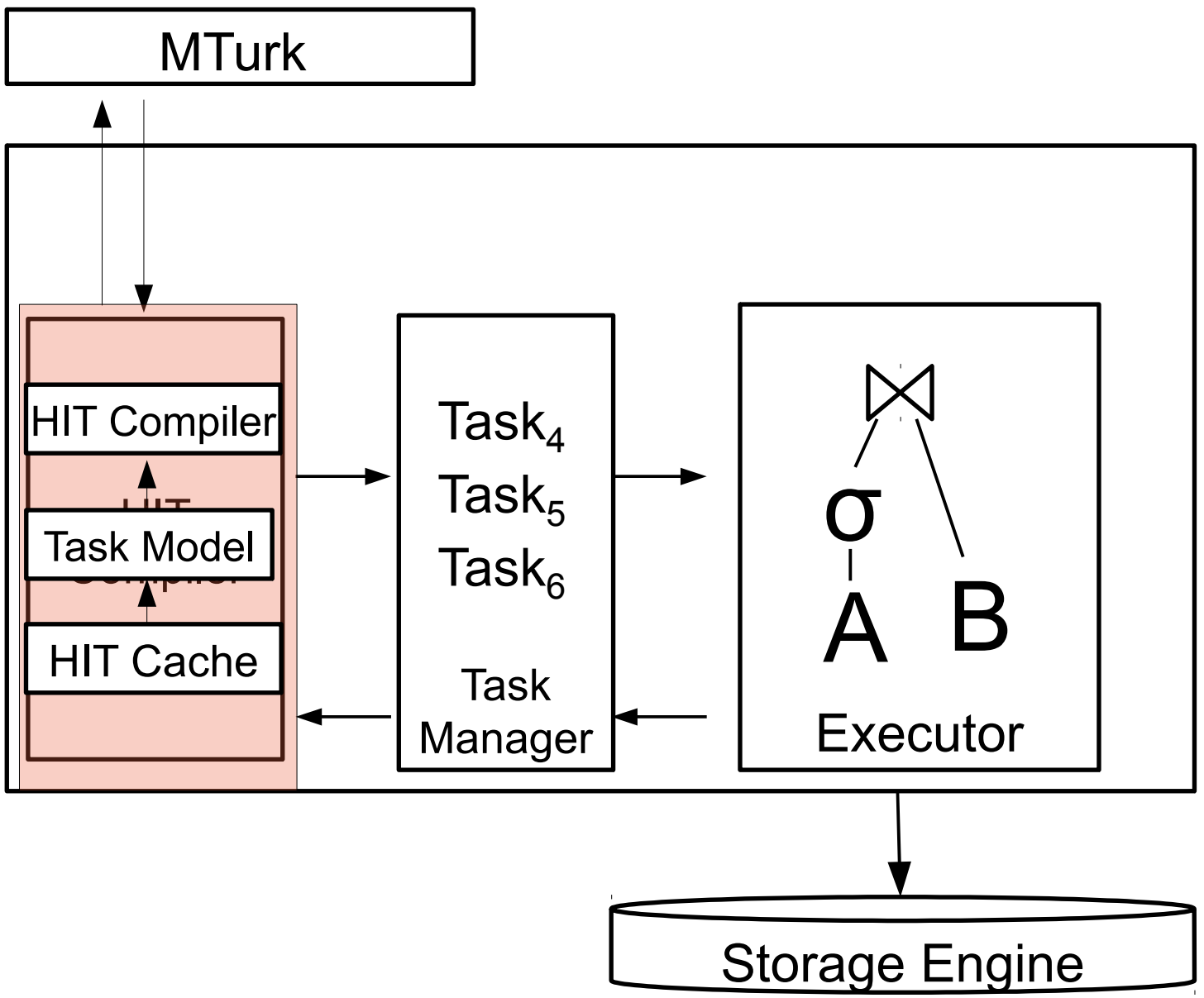


Batch Tuples and Operators

Operator Optimizations

Selectivity Injection

HIT Avoidance

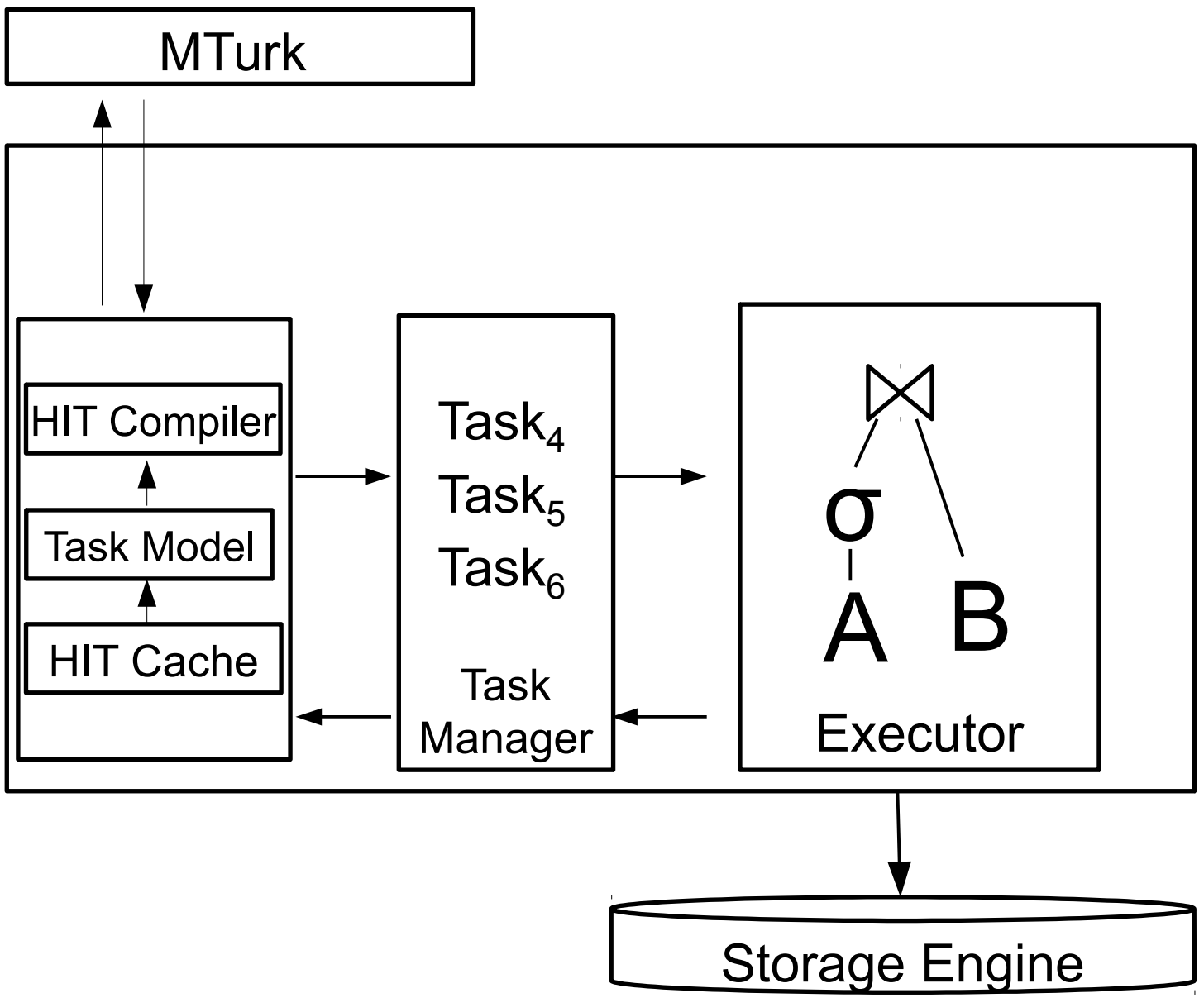


Batch Tuples and Operators

Operator Optimizations

Selectivity Injection

HIT Avoidance



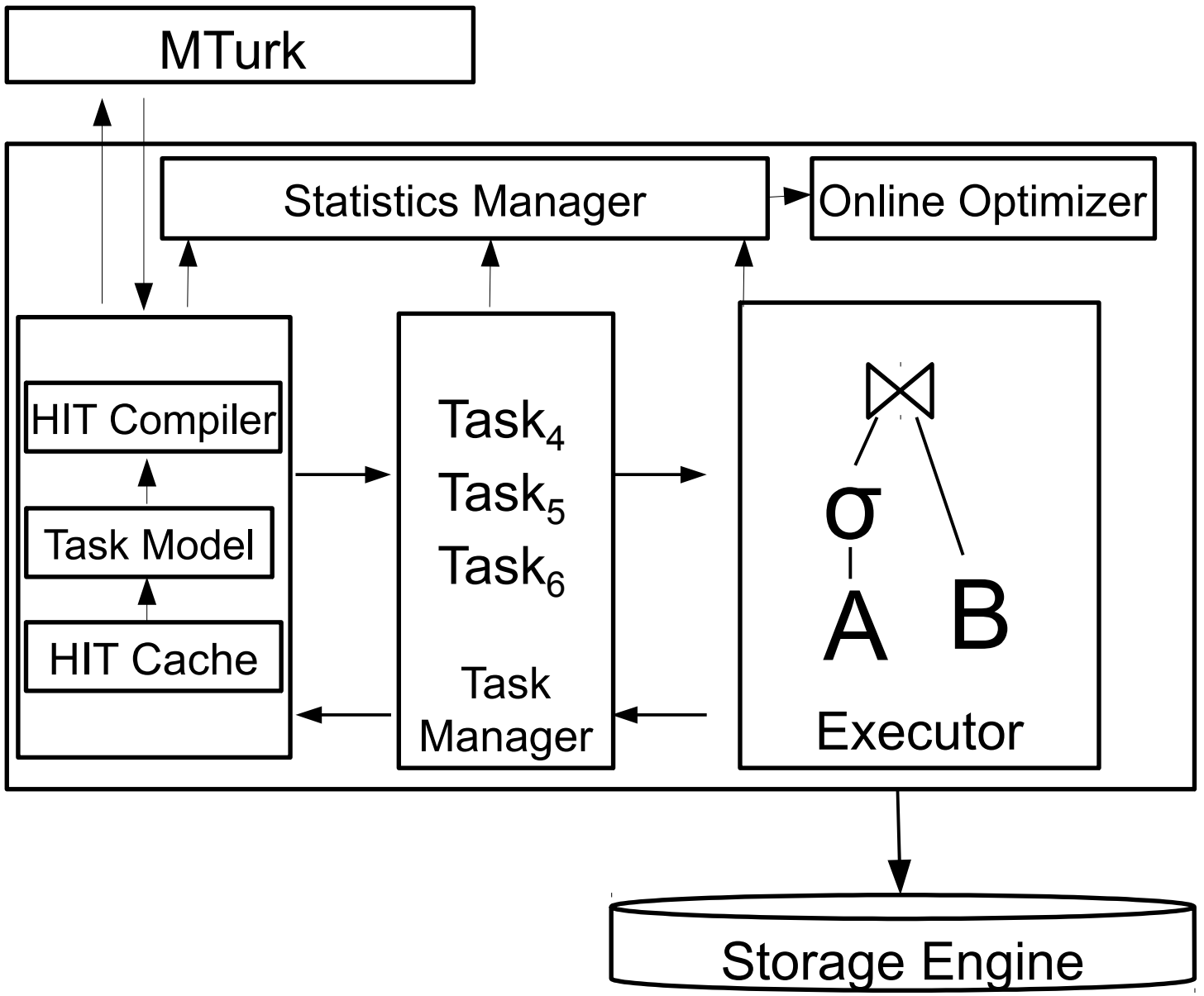
Batch Tuples and Operators

Operator Optimizations

Selectivity Injection

HIT Avoidance

Online Optimizations



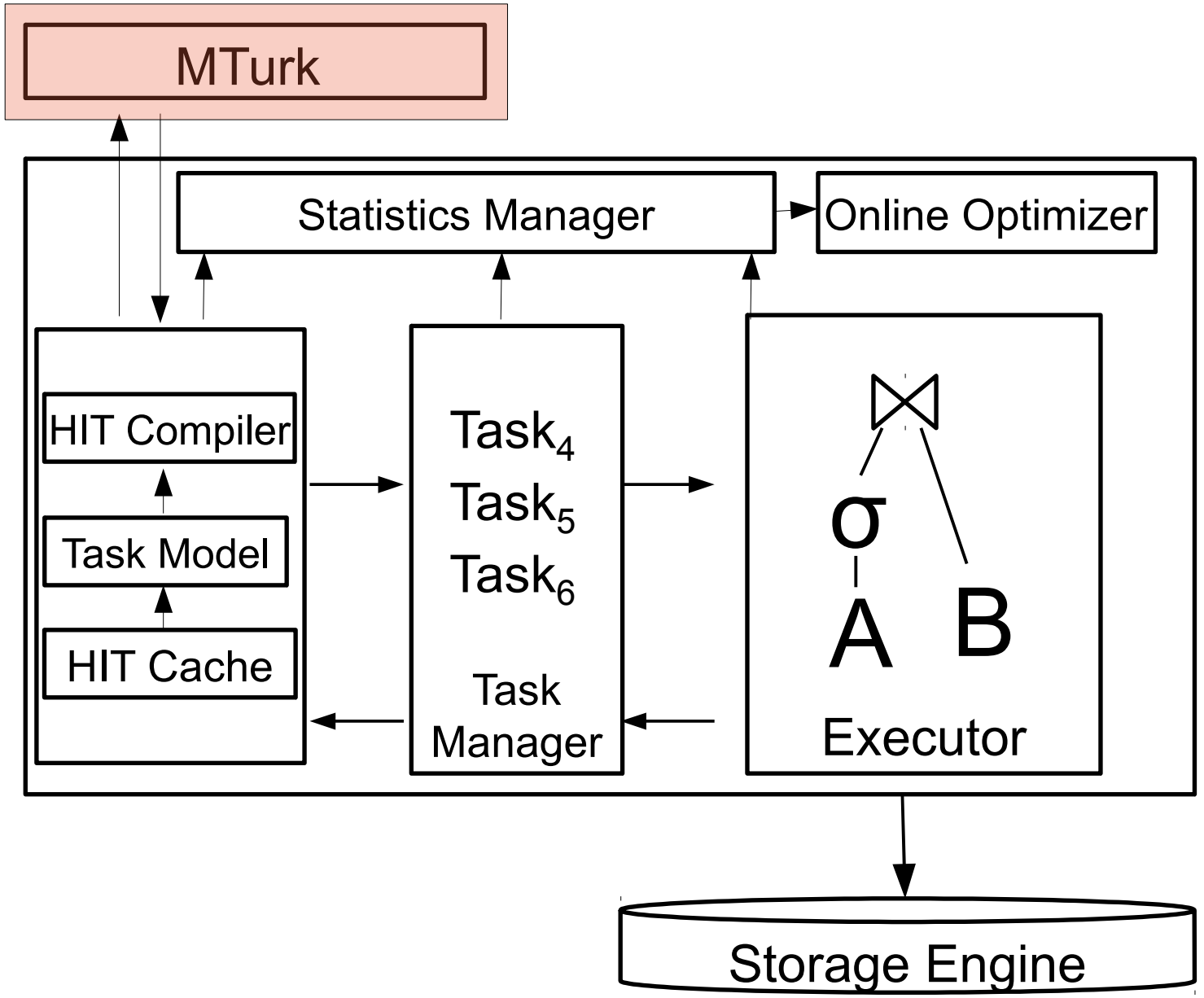
Batch Tuples and Operators

Operator Optimizations

Selectivity Injection

HIT Avoidance

Online Optimizations



Batch Tuples and Operators

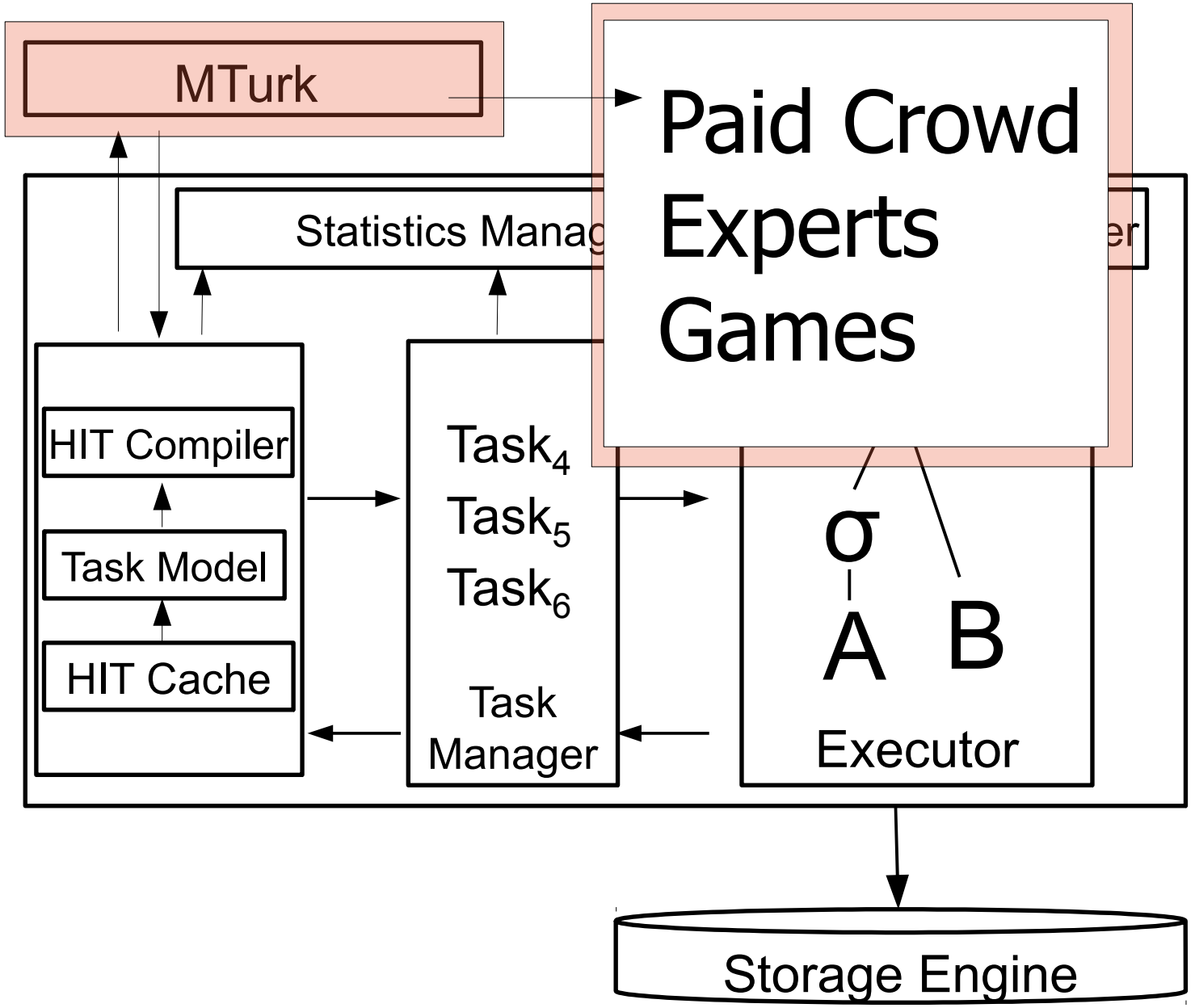
Operator Optimizations

Selectivity Injection

HIT Avoidance

Online Optimizations

Human Computation Platforms



Batch Tuples and Operators

Operator Optimizations

Selectivity Injection

HIT Avoidance

Online Optimizations

Human Computation Platforms