Data Management Infrastructure

CHIDATA

Things That Are Different

Bulk Storage System: Stores raw data reliably at scale.

Amazon S3

Database: Organizes structured data allowing for fast information retrieval

Postgres, Oracle, Amazon Redshift

Data Analytics Framework: Perform data manipulation or analytics at scale

Apache Spark

Things That Are Different

Bulk Storage System: Stores raw data reliably at scale.

Database: Organizes structured data allowing for fast information retrieval

Data Analytics Framework: Perform data manipulation or analytics at scale

Bulk Storage Systems

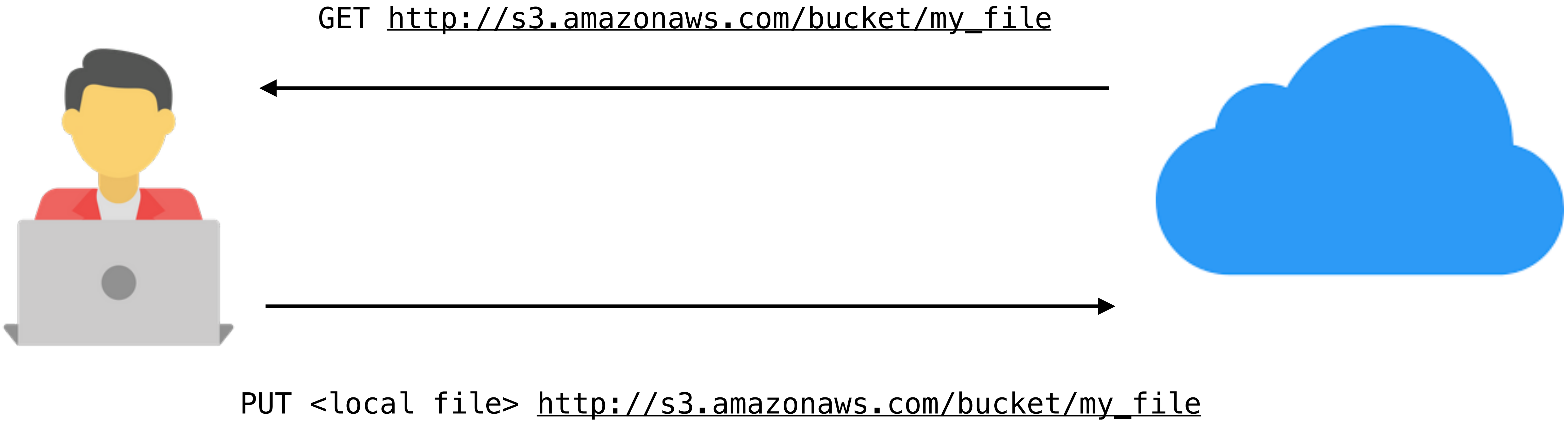
- Store and retrieve large amounts of raw data files (or blobs)
- Pay-as-you-go cloud services
- Cheap cost per byte

amazon S3

Google Cloud Storage

Azure Blob Storage

Bulk Storage Systems

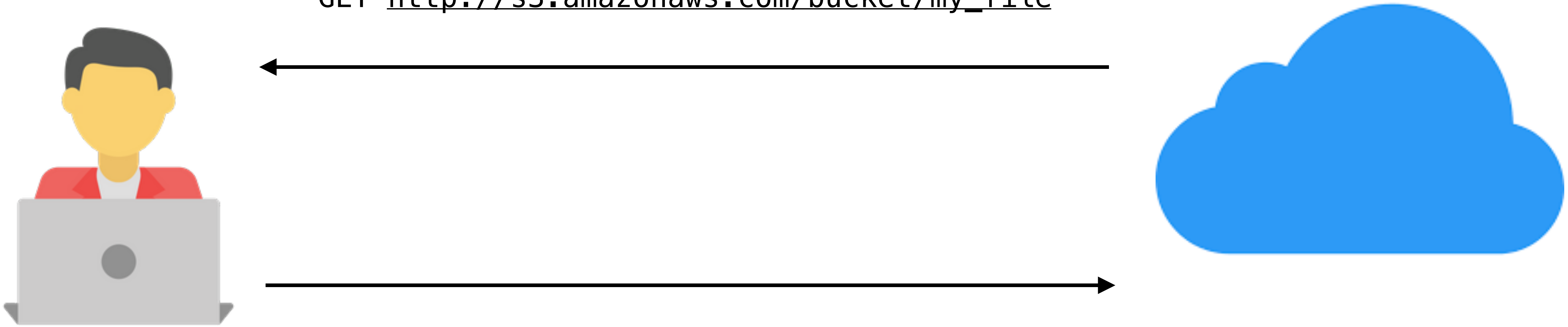


GET http://s3.amazonaws.com/bucket/my_file

PUT <local file> http://s3.amazonaws.com/bucket/my_file

- Retrieval is solely by file name
- No content-based search features

Bulk Storage Systems

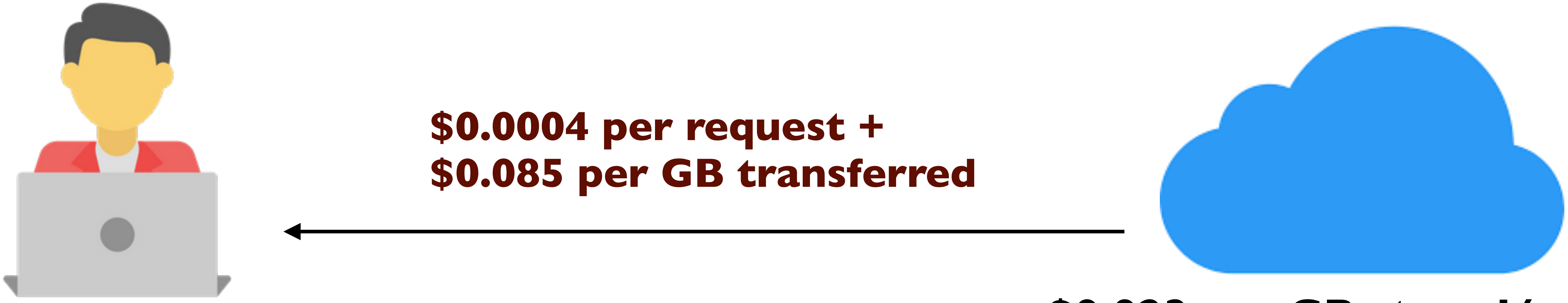


GET http://s3.amazonaws.com/bucket/my_file

PUT <local file> http://s3.amazonaws.com/bucket/my_file

- **Scalability**: Performance doesn’t degrade with more data
- **Durability**: Data is (basically) never lost
- **Availability**: Data is available for access almost all of the time

Pricing (Amazon)



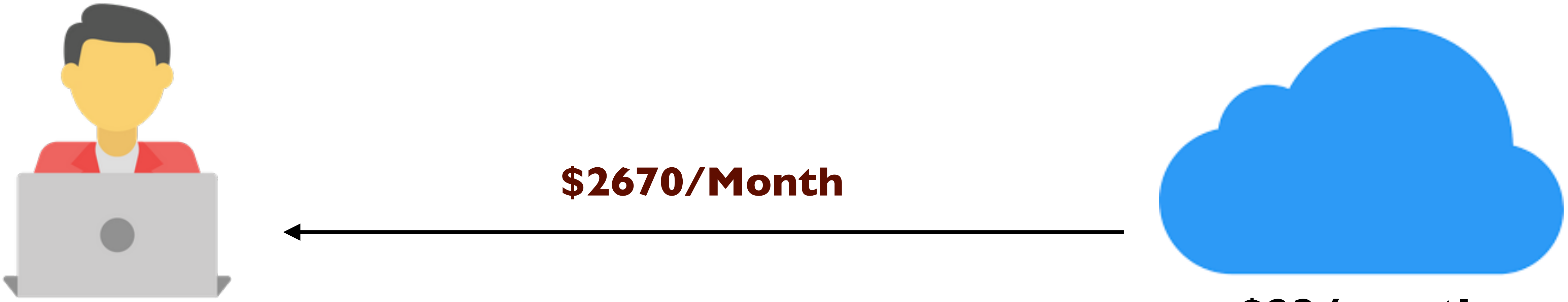
$0.0004 per request +
$0.085 per GB transferred

$0.023 per GB stored/month

Costs you almost nothing to keep large data on the cloud, but using it costs a lot more!

Pricing (Amazon)

Video sharing website that stores 1TB of data, visited by 10000 people each day who watch 100 MB of video.



$2670/Month

$23/month

CHIDATA

Things That Are Different

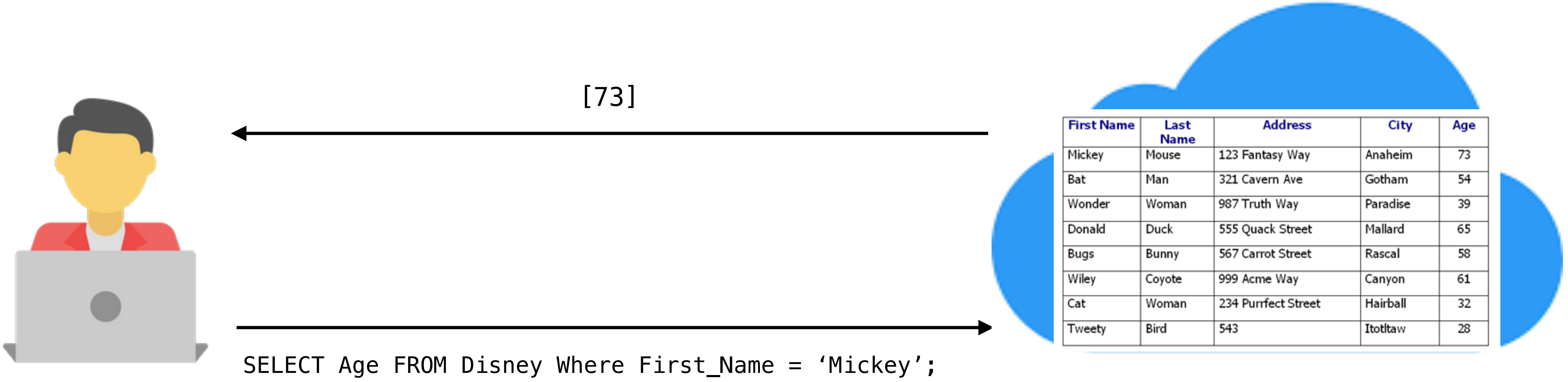
Bulk Storage System: Stores raw data reliably at scale.

Database: Organizes structured data allowing for fast information retrieval

Data Analytics Framework: Perform data manipulation or analytics at scale

Database Systems

- Store and retrieve structured data selectively.
- SQL user-interface



```
[73]
```

```
SELECT Age FROM Disney Where First_Name = ‘Mickey’;
```

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

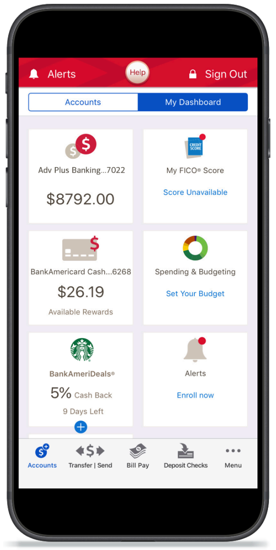
Database Systems

Transaction Processing v.s. Analytics Processing

- Lot’s of concurrent accesses
- Selective modification of the data

Websites

Banking



Adv Plus Banking...7022
$8792.00

My FICO® Score
Score Unavailable

BankAmericard Cash...6268
$26.19
Available Rewards

Spending & Budgeting
Set Your Budget

BankAmeriDeals®
5% Cash Back
9 Days Left

Alerts
Enroll now

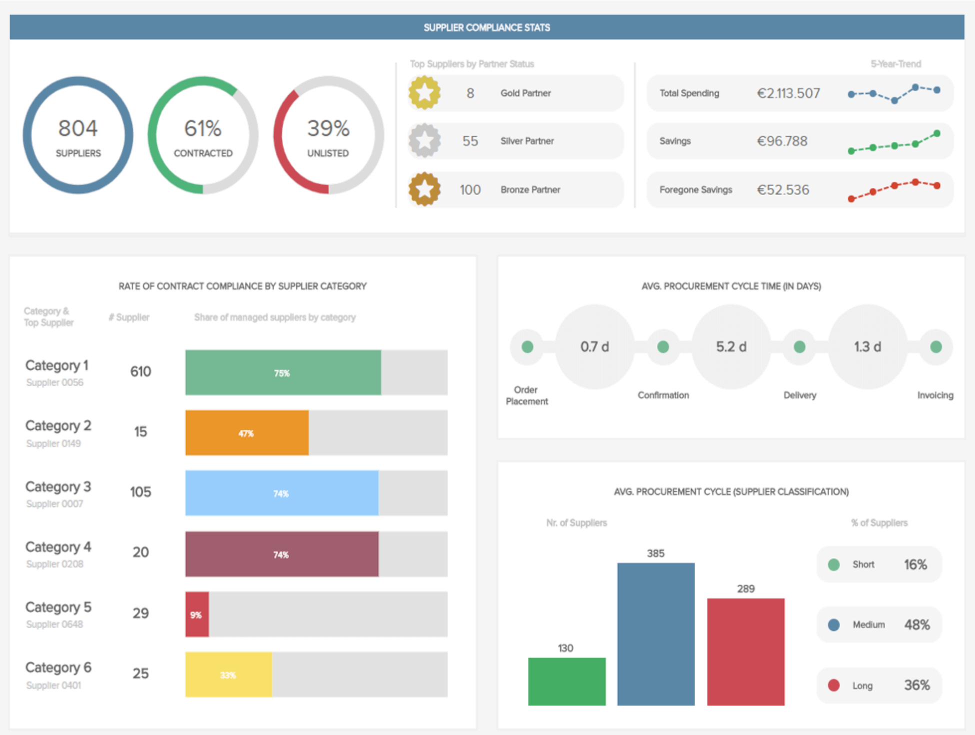
CHIDATA

Database Systems

Transaction Processing v.s. **Analytics Processing**

- Static data or append-only data
- Pre-computed results (or views)

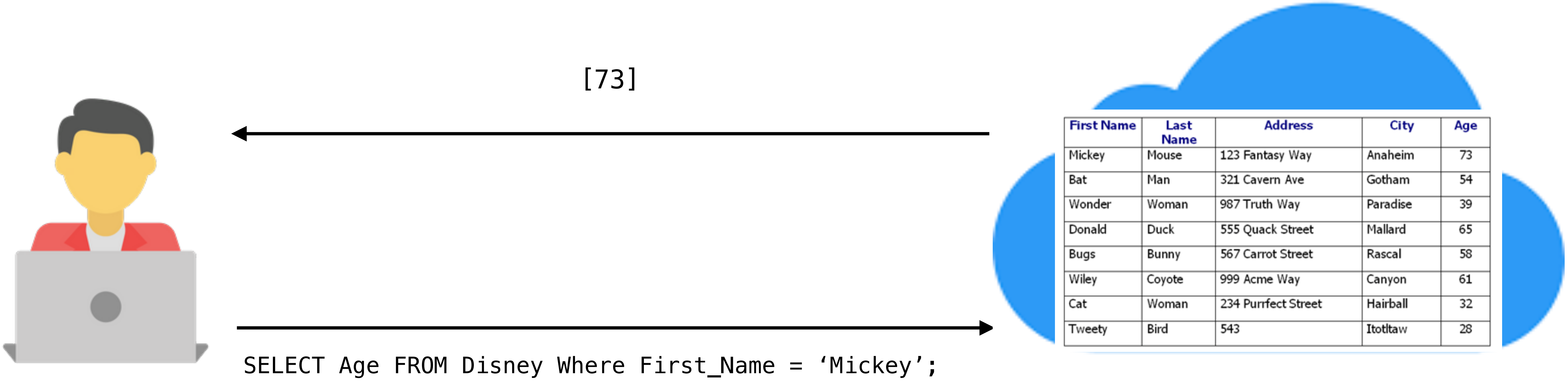
Dashboards


SUPPLIER COMPLIANCE STATS
804
SUPPLIERS
61%
CONTRACTED
39%
UNLISTED
Top Suppliers by Partner Status
8 Gold Partner
55 Silver Partner
100 Bronze Partner
5-Year-Trend
Total Spending €2.113.507
Savings €96.788
Foregone Savings €52.536
RATE OF CONTRACT COMPLIANCE BY SUPPLIER CATEGORY
Category & Top Supplier
# Supplier
Share of managed suppliers by category
Category 1 Supplier 0056 610 75%
Category 2 Supplier 0149 15 47%
Category 3 Supplier 0007 105 74%
Category 4 Supplier 0208 20 74%
Category 5 Supplier 0648 29 9%
Category 6 Supplier 0401 25 33%
AVG. PROCUREMENT CYCLE TIME (IN DAYS)
0.7 d
5.2 d
1.3 d
Order Placement
Confirmation
Delivery
Invoicing
AVG. PROCUREMENT CYCLE (SUPPLIER CLASSIFICATION)
Nr. of Suppliers
130
385
289
% of Suppliers
Short 16%
Medium 48%
Long 36%

Reports

Database Systems

- Only work for structure data
- Much more expensive than using a bulk storage system! (about 20x)



[73]

```
SELECT Age FROM Disney Where First_Name = ‘Mickey’;
```

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Things That Are Different

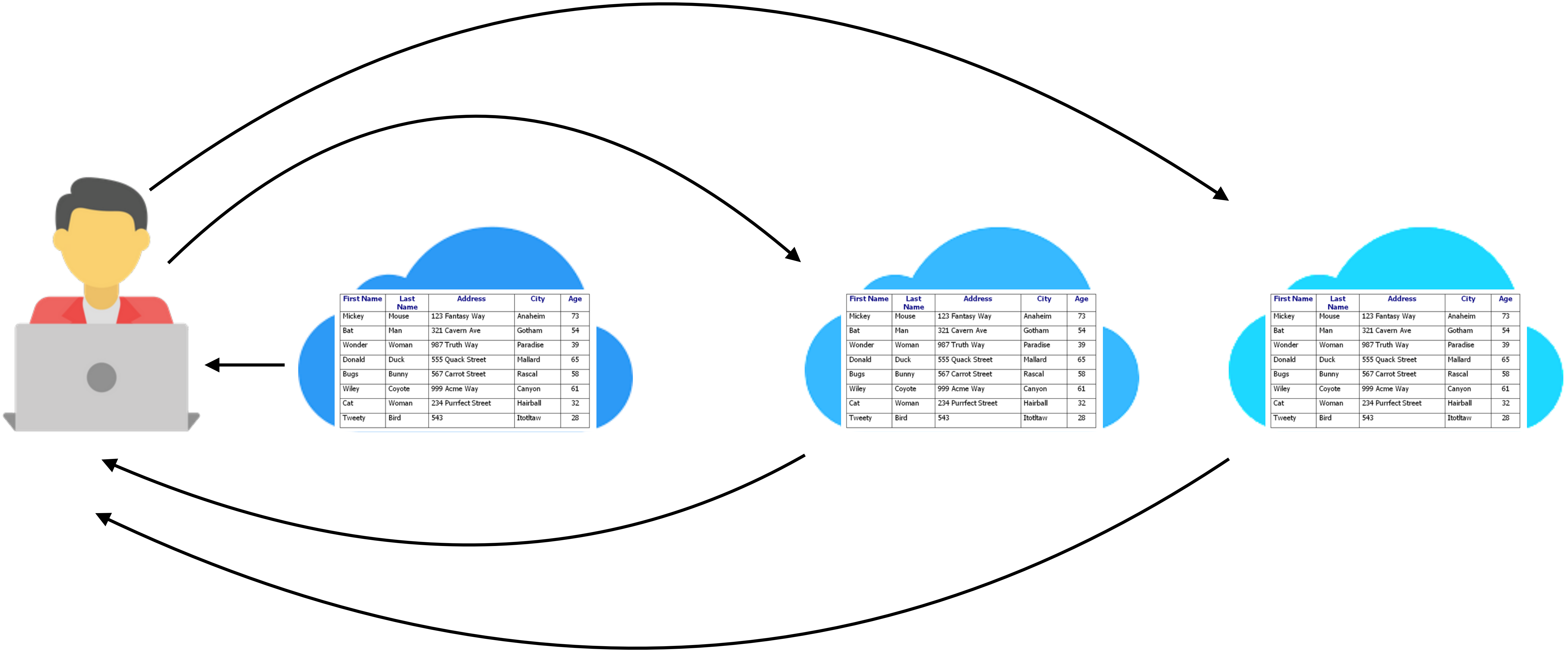
Bulk Storage System: Stores raw data reliably at scale.

Database: Organizes structured data allowing for fast information retrieval

Data Analytics Framework: Perform data manipulation or analytics at scale

Analytics Beyond SQL

Many analytics tasks are iterative and multi-stage



First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Analytics Beyond SQL

Example. Fix names before analysis

Id	Name	Salary	Age
1	John Davis	123232	31
2	Carlton, Jenny	439921	43
3	Stillman, Pat	553229	32
4	Donny Gent	39832	23

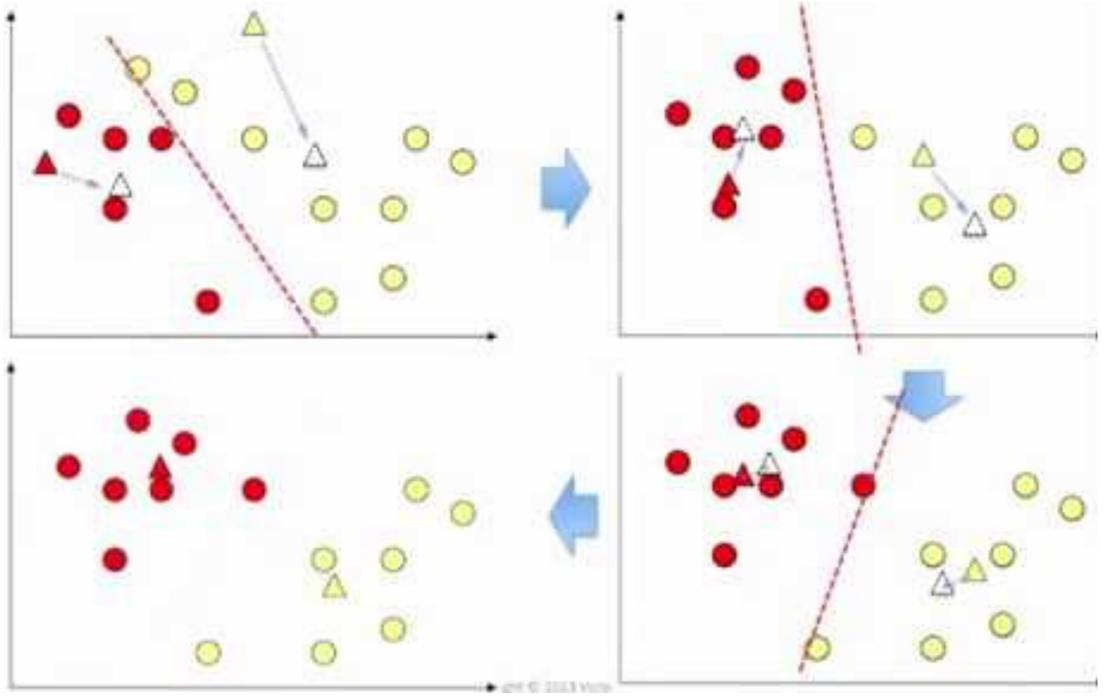
↓

Id	First Name	Last Name	Salary	Age
1	John	Davis	123232	31
2	Jenny	Carlton	439921	43
3	Pat	Stillman	553229	32
4	Donny	Gent	39832	23

Analytics Beyond SQL

Example. Data Clustering

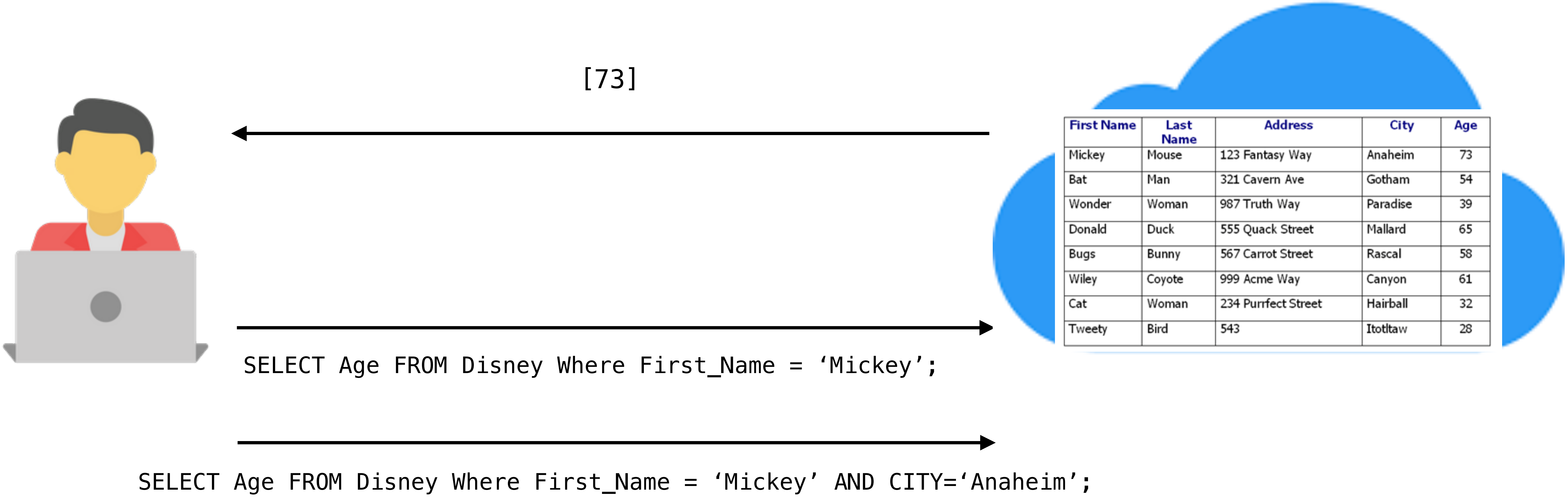
Id	Name	Salary	Age
1	John Davis	123232	31
2	Carlton, Jenny	439921	43
3	Stillman, Pat	553229	32
4	Donny Gent	39832	23



Database Systems

Queries are generally independent from each other

Little to no result caching between queries.



[73]

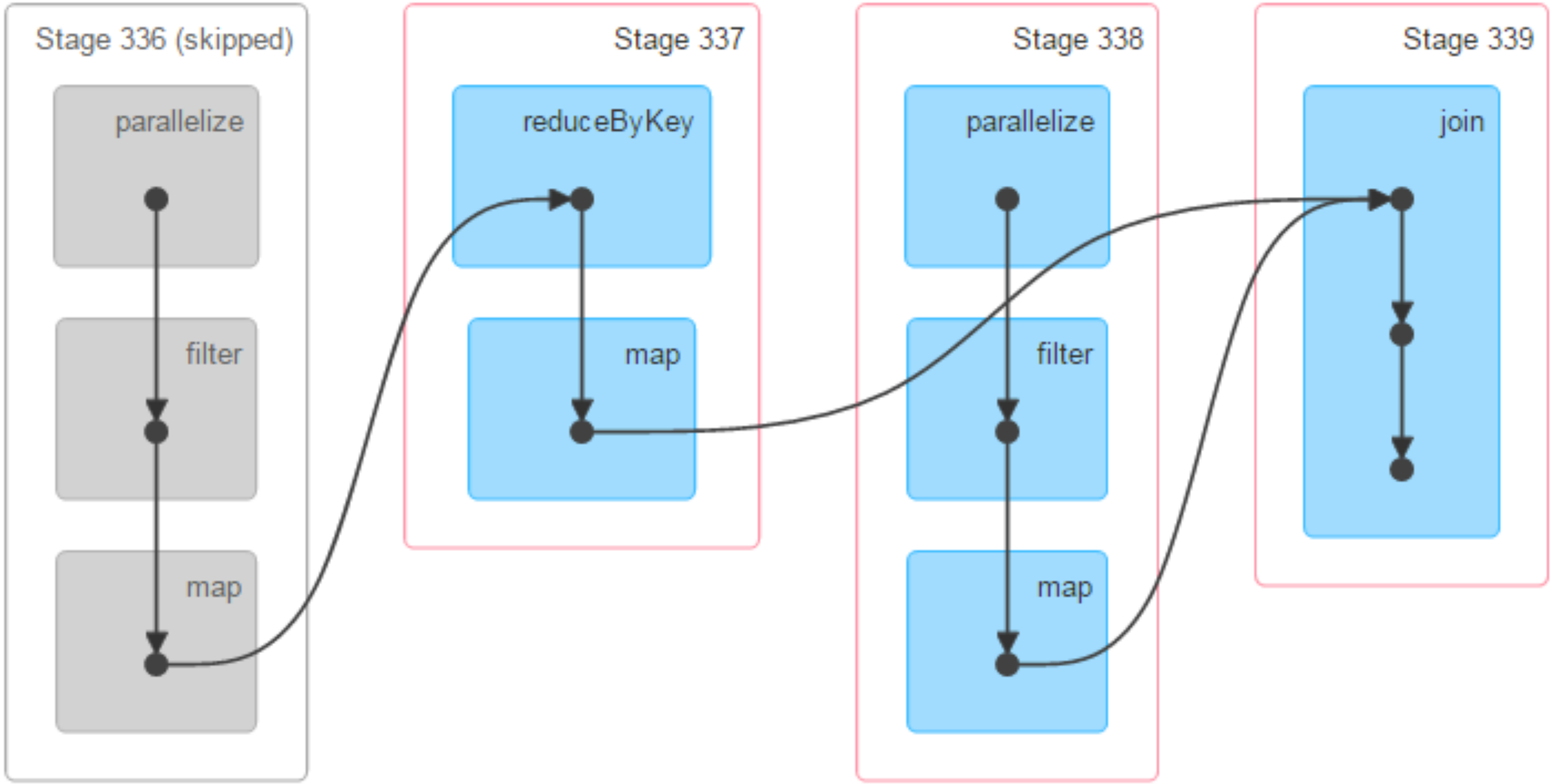
```
SELECT Age FROM Disney Where First_Name = ‘Mickey’;
```

```
SELECT Age FROM Disney Where First_Name = ‘Mickey’ AND CITY=‘Anaheim’;
```

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Apache Spark

Data analytics framework optimized for iterative/stage-wise analytics



Apache Spark

Find all log entries with an error

```
11-Nov mysql ERROR Out-of-memory error
12-Nov php WARN php 5.2 is out of date
…
```

```
val lines = sc.textFile(“log.txt")  //input

val errors = lines.filter(_.startsWith("ERROR"))  //filter

val messages = errors.map(_.split("\t")).map(r => r(1))  //map
messages.cache() //materialize results

// count 1
messages.filter(_.contains("mysql")).count()
// count 2
messages.filter(_.contains("php")).count()
```

Things That Are Different

Bulk Storage System: Stores raw data reliably at scale.

Amazon S3

Database: Organizes structured data allowing for fast information retrieval

Postgres, Oracle, Amazon Redshift

Data Analytics Framework: Perform data manipulation or analytics at scale

Apache Spark

Data Governance

Organizational policies determining how data can be stored and used.

Privacy: How can user data be used, shared, and who can see it

Data Provenance: How a result is derived

Privacy

Policy stipulations on allowed data processing

FERPA - Educational data

HIPAA - Medical data

GDPR - General data privacy law in the EU zone

Today's laws essentially disallow "linkage" of identifiable or sensitive information.

Privacy

Policy stipulations on allowed data processing

FERPA - Educational data

HIPAA - Medical data

GDPR - General data privacy law in the EU zone

Today’s laws essentially disallow “linkage” of identifiable or sensitive information by third-parties.

Example

id	Name	Major	GPA
964682	John Davis	BIO	3.71
198227	Travis Park	LIT	2.95
443521	Anna Black	PHY	3.91
440921	Erin Thomas	BIO	3.02
•••	…	…	…

Example

Should not be able to link Student Id or Name to GPA without outside information.

Student id	Name	Major	GPA
964682	John Davis	BIO	3.71
198227	Travis Park	LIT	2.95
443521	Anna Black	PHY	3.91
440921	Erin Thomas	BIO	3.02
…	…	…	…

One Solution

Pseudo Anonymous	[redacted]	Major	GPA
1	[redacted]	BIO	3.71
2	[redacted]	LIT	2.95
3	[redacted]	PHY	3.91
4	[redacted]	BIO	3.02
…	[redacted]	…	…

Privacy Implications

Bulk Storage System: Stores raw data reliably at scale.

Hard to determine if data are linked

Database: Organizes structured data allowing for fast information retrieval

Easier because the data are more structured

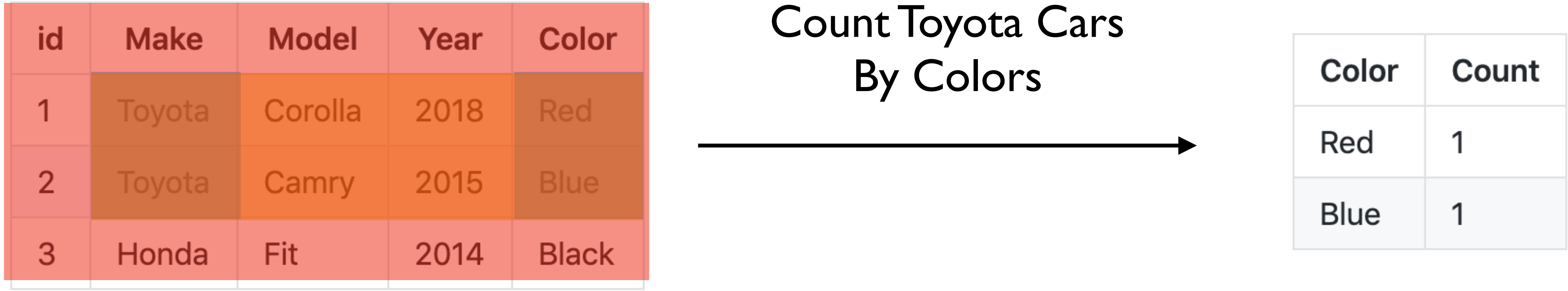
Data Analytics Framework: Perform data manipulation or analytics at scale

Somewhere in between

Data Provenance

What data contributed to a result

Provenance can be tracked at different granularities



id	Make	Model	Year	Color
1	Toyota	Corolla	2018	Red
2	Toyota	Camry	2015	Blue
3	Honda	Fit	2014	Black

Count Toyota Cars
By Colors

Color	Count
Red	1
Blue	1

CHIDATA

Provenance Implications

Bulk Storage System: Stores raw data reliably at scale.

Can only track at the level of files

Database: Organizes structured data allowing for fast information retrieval

Can track at the level of tuples (in principle…)

Data Analytics Framework: Perform data manipulation or analytics at scale

Can track at the level of data partitions