Scalable, Flexible Data Privacy in the Cloud

motorola
a Google company

"The views expressed in this presentation are generic in nature and do not reflect any endorsement by Motorola Mobility LLC."

Agenda Outline

- Data Privacy in the Cloud
- Security vs Privacy
- What is Anonymization?
- Anonymization Techniques
- Hadoop Anonymization Toolkit (HAT)
- Conclusions
- Q&A

Data Privacy in the Cloud (Opportunities and Challenges)

Privacy in the Cloud (Opportunities)

- Volume and granularity of data collection is exploding
- Data is available in public and private domains
- Applications that provide targeted, personalized and contextual experiences to customers have tremendous business value and competitive advantage

These opportunities can be realized if we solve privacy responsibly.

Privacy in the Cloud (Challenges)

- Data sharing is important but has to be done without revealing sensitive information

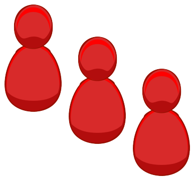
- Existing big data systems have PII vulnerabilities and pitfalls

- 87% of the U.S. population can be uniquely identified using gender, date of birth, and zip code attributes [3]

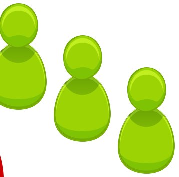
A big data solution is required to address these challenges in conjunction with meeting business needs

Security vs Privacy Protection

Authentication & Authorization safeguard against malicious data access



Personally identifiable information can still be at risk behind the firewall

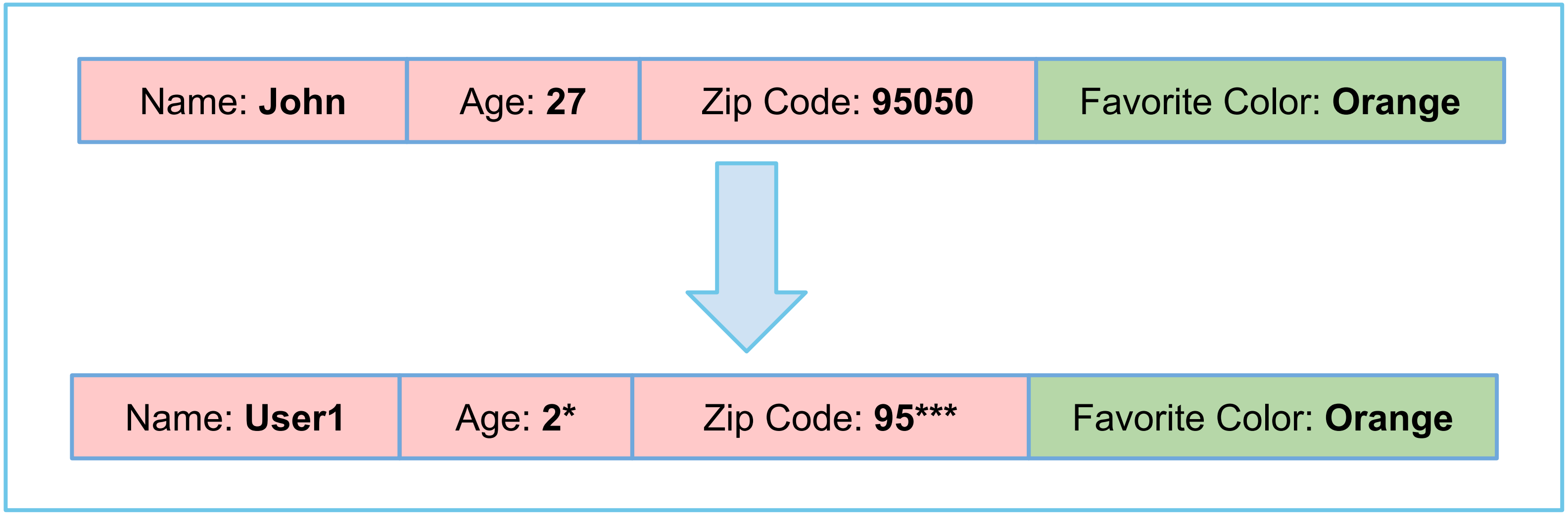


How to protect data privacy while keeping data useful to derive insights?

Anonymization

What is Anonymization?

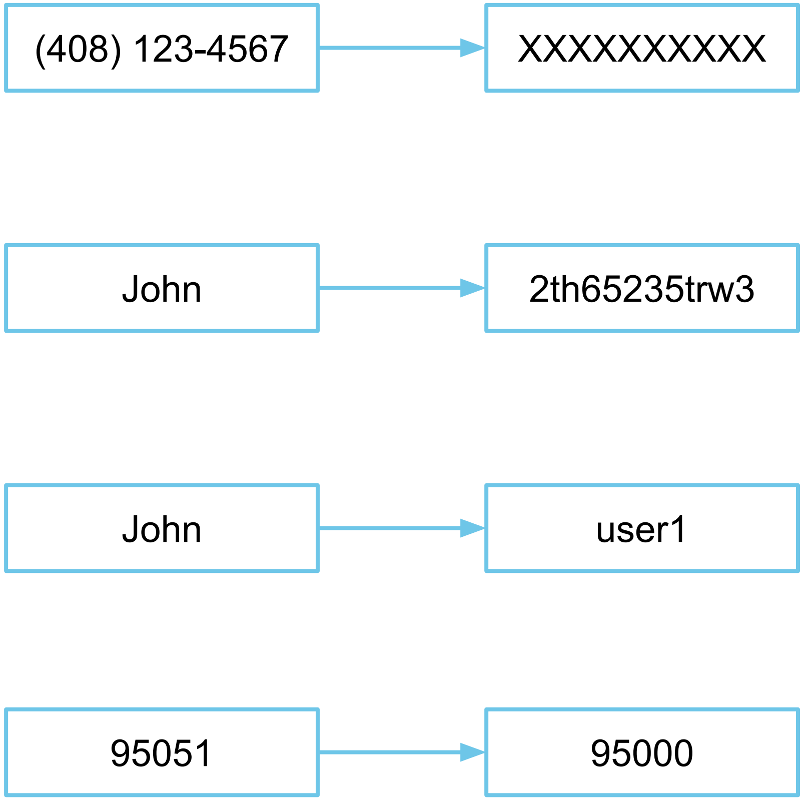
- Confidential data / sensitive information can be obscured in ways that maintains privacy while preserving the ability to derive useful insights*



Name: **John**	Age: **27**	Zip Code: **95050**	Favorite Color: **Orange**
Name: **User1**	Age: **2***	Zip Code: **95*****	Favorite Color: **Orange**

* Refer to [4] for definition and further discussion.

Anonymization Techniques



- Hiding

(408) 123-4567 → XXXXXXXXXX

- Hashing

John → 2th65235trw3

- Mapping

John → user1

- Editing

95051 → 95000

Hadoop Anonymization Toolkit (HAT)

- There is a gap in the existing big data ecosystem for a generic and extensible anonymization solution

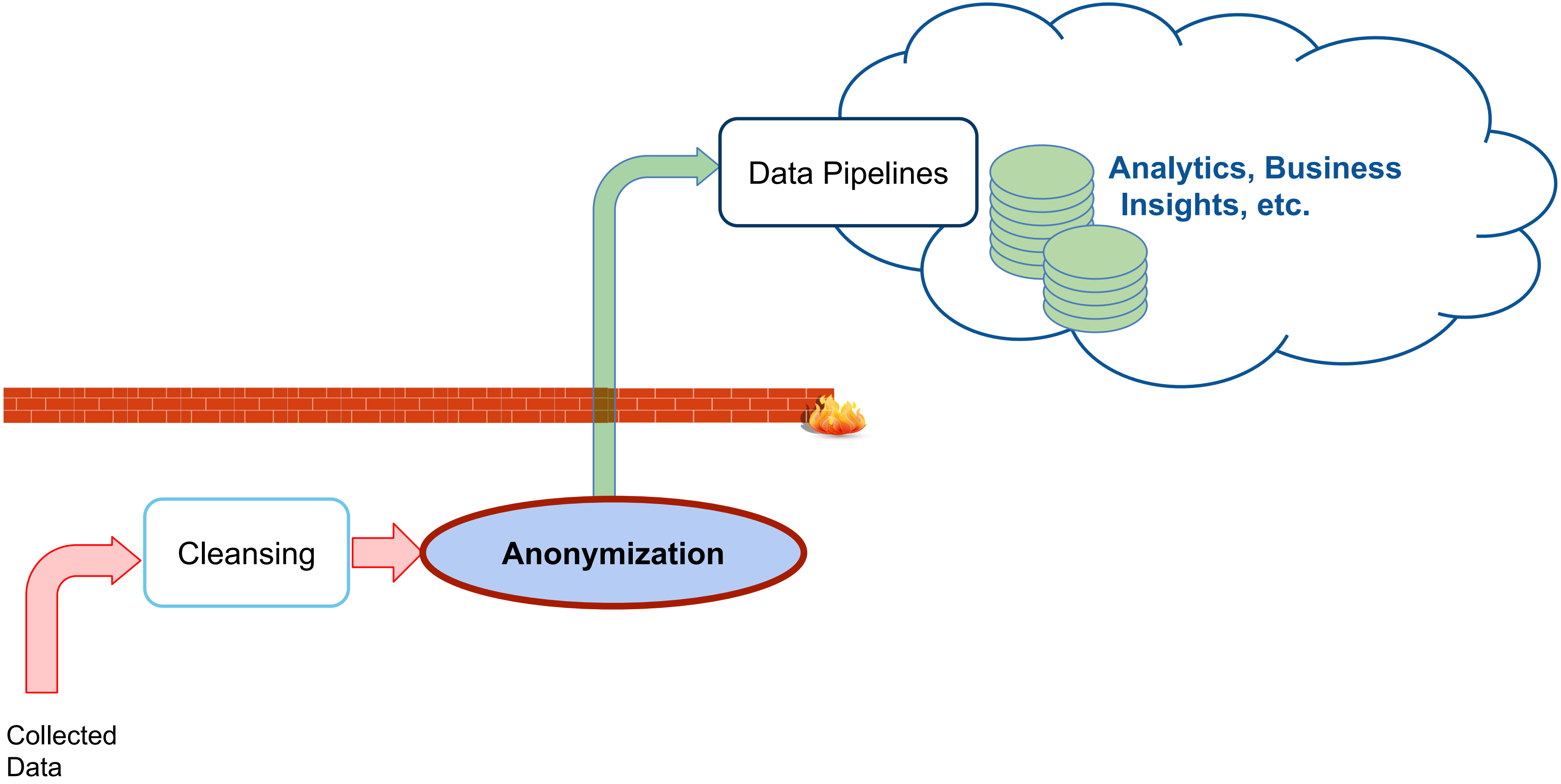
- A need for a scalable solution that can help organizations solve their big data anonymization needs

HAT is a solution that addresses these needs.

HAT Main Objectives

- Configuration driven framework for simple and flexible composition of anonymization jobs
- Scalable to data size outbursts or fluctuations
- Adapts to schema changes and evolution
- Existing anonymization primitives can be extended or new ones added.

Anonymization in the Data Continuum



Data Pipelines

Analytics, Business Insights, etc.

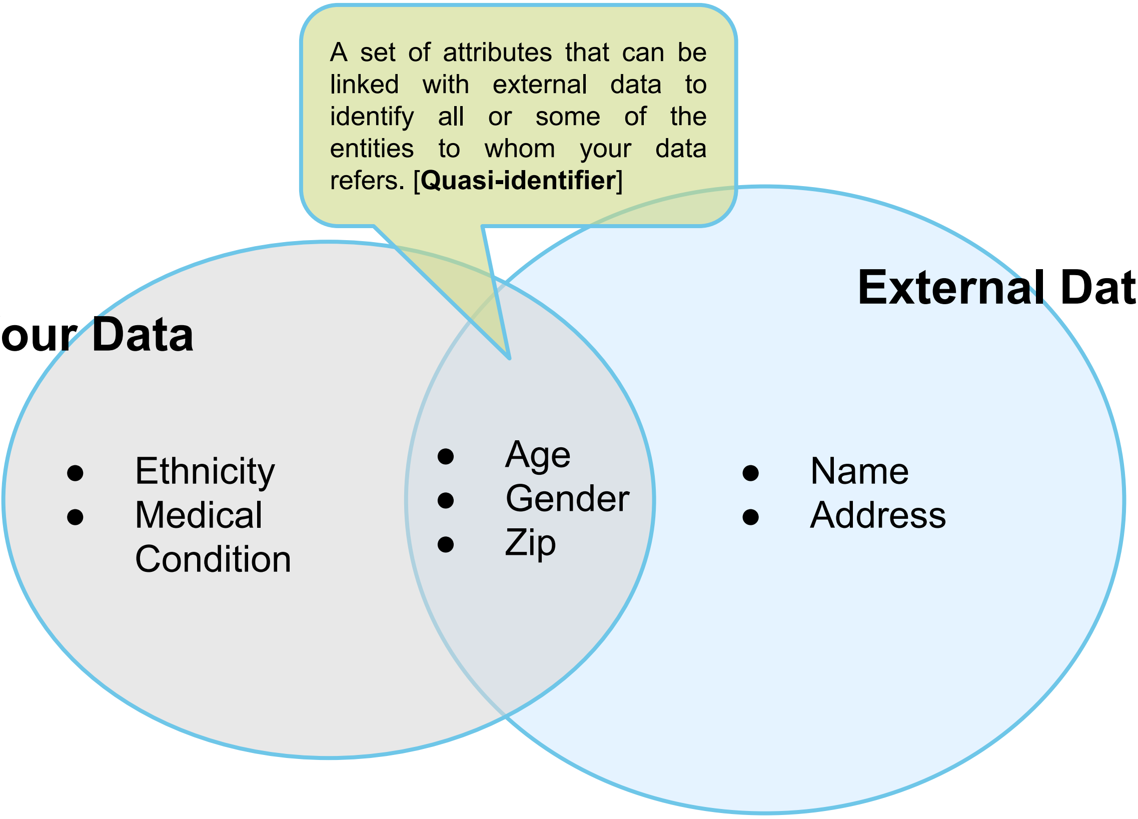
Cleansing

Anonymization

Collected Data

Modeling for Robust Anonymization

A Health Care Example



A set of attributes that can be linked with external data to identify all or some of the entities to whom your data refers. [**Quasi-identifier**]

Your Data

- Ethnicity
- Medical Condition

- Age
- Gender
- Zip

External Data

- Name
- Address

Identification by linking

K-Anonymity Concepts

- The goal is to make each record indistinguishable from a defined number (k) of other records

- In a k-anonymized dataset, each record is indistinguishable from at least k −1 other records with respect to certain “identifying” attributes

- In general the higher the value of k, the more data privacy is achieved

Health Care Example (cont.)

4 - Anonymous dataset

Age	Gender	Zip Code	Condition
2*	*	95***	Heart Disease
2*	*	95***	Viral Infection
2*	*	95***	Heart Disease
2*	*	95***	Viral Infection
3*	*	94***	Viral Infection
3*	*	94***	Heart Disease
3*	*	94***	Cancer
3*	*	94***	Viral Infection
4*	*	95***	Cancer
4*	*	95***	Cancer
4*	*	95***	Cancer
4*	*	95***	Cancer

Background Knowledge Attack

Homogeneity Attack

L-Diversity

- k-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute.

- In addition to k-anonymity, the data should also ensure “diversity” – all tuples that share the same values of their **quasi-identifiers** should have diverse values for their sensitive attributes.

- The main idea behind ℓ-diversity is that the values of the sensitive attributes are well-represented in each group to protect data privacy

Anonymization Measures

K-Anonymity

Age	Gender	Zip Code	Condition
3*	*	94***	Viral Infection
3*	*	94***	Heart Disease
3*	*	94***	Cancer
3*	*	94***	Viral Infection

L-Diversity

A combination of these measures is usually required for robust anonymization of data

HAT I/O Schemas and Configuration

Input

```
{
  "name": "John",
  "zip_code": "95050",
  "age": "27",
  "favorite_color": "Orange"
}
```

Configuration

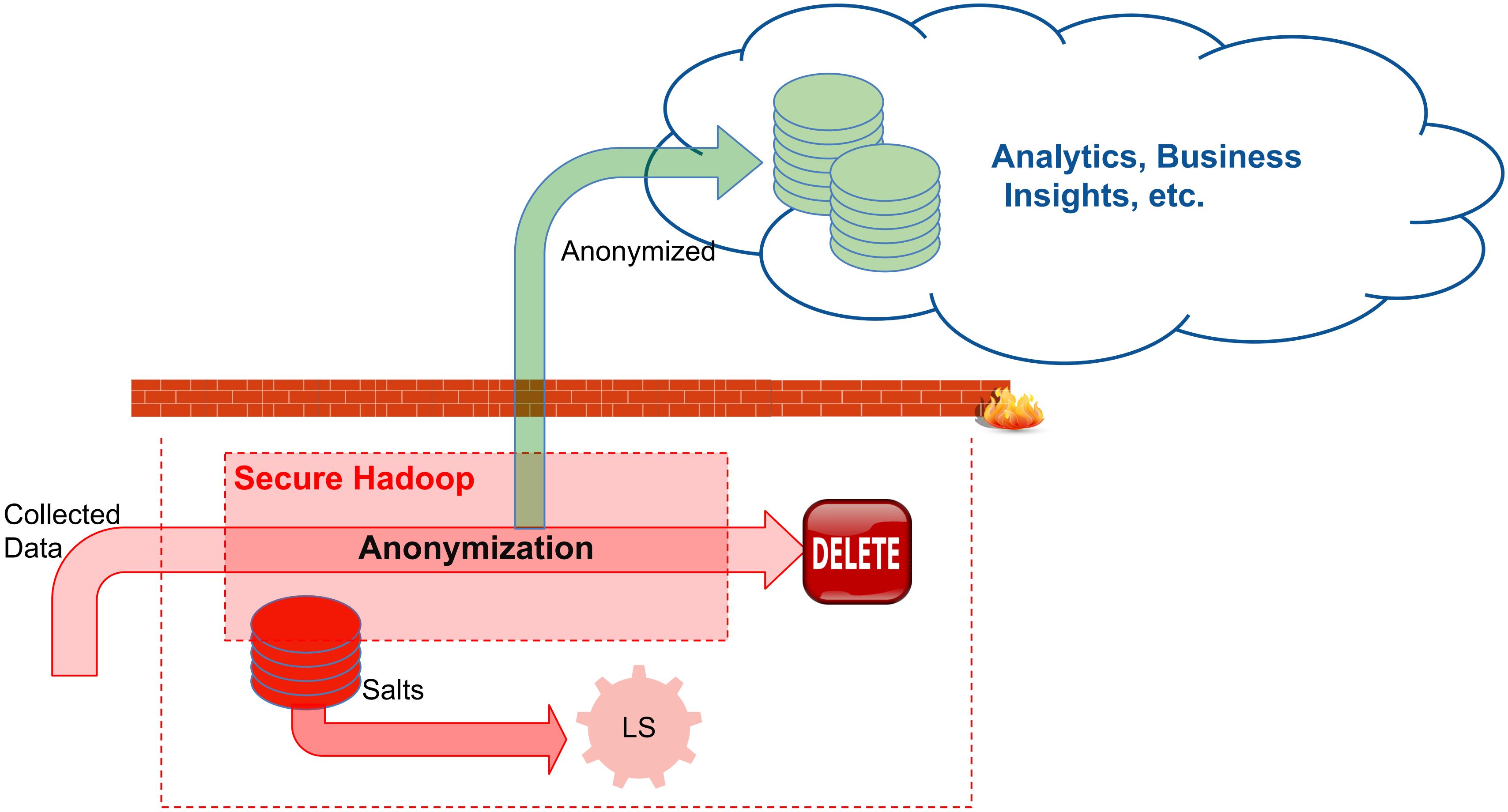
```
{
   "obfuscate": [
      {
         "name": {
            "lookup": "$JSON.get(\"name\")",
            "update": "$JSON.put(\"name\",\"$OBFUSCATED\")",
            "obfuscation-type": "UNIQUE_PER_VALUE"
         }
      },
      ...
   ]
}
```

Output

```
{
  "name": "user1",
  "zip_code": "95***",
  "age": "2*",
  "favorite_color": "Orange"
}
```

- JSON input and output
- JSON expressions → what fields to obfuscate
- Obfuscation types
- Support for repeated elements

Anonymization Data Flow



Hashing Salts Table

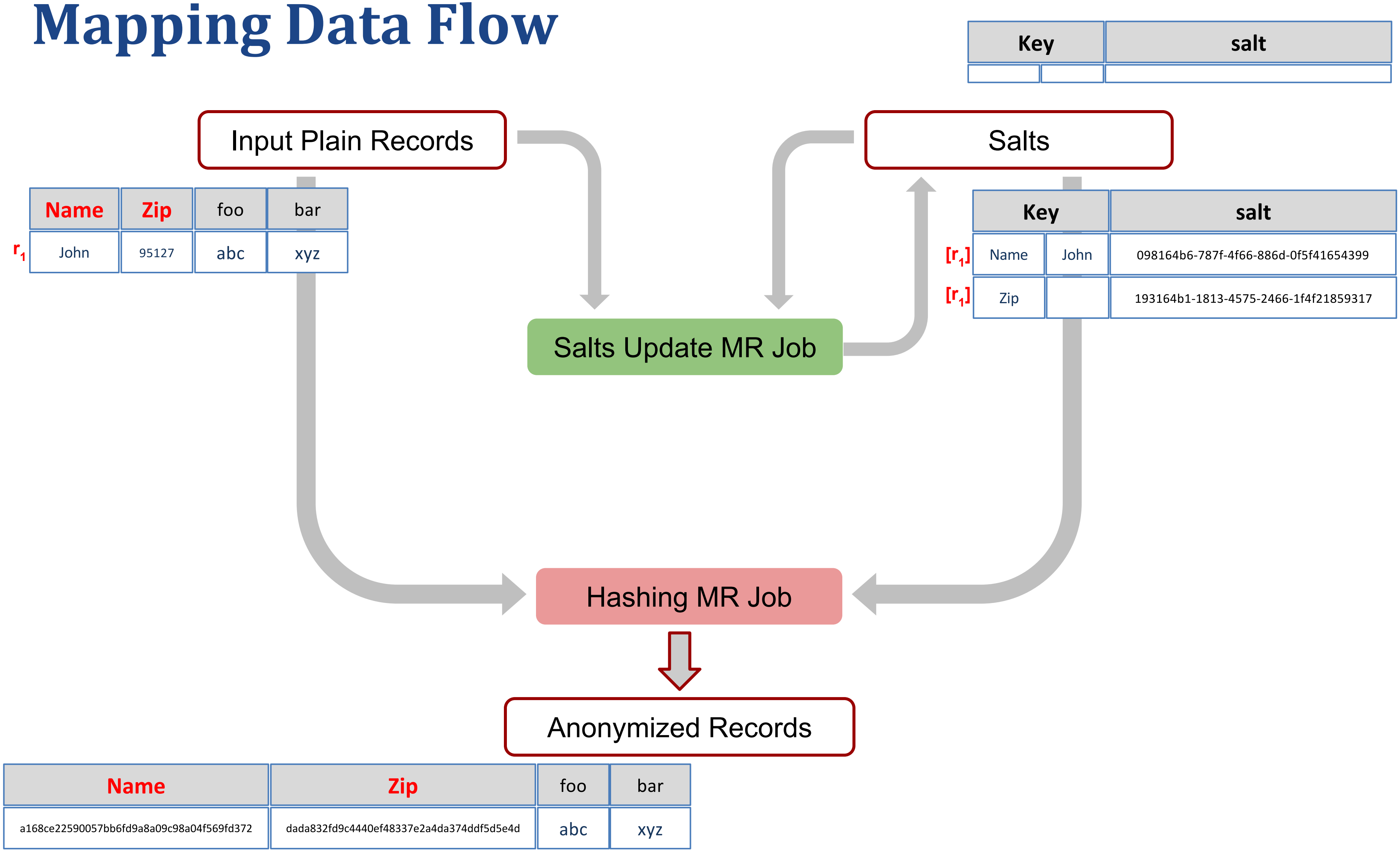
- Unique per value

Key		salt
Name	John Smith	098164b6-787f-4f66-886d-0f5f41654399
Zip Code		193164b1-1813-4575-2466-1f4f21859317

- Unique per type

- Groups of elements sharing same salt

Mapping Data Flow



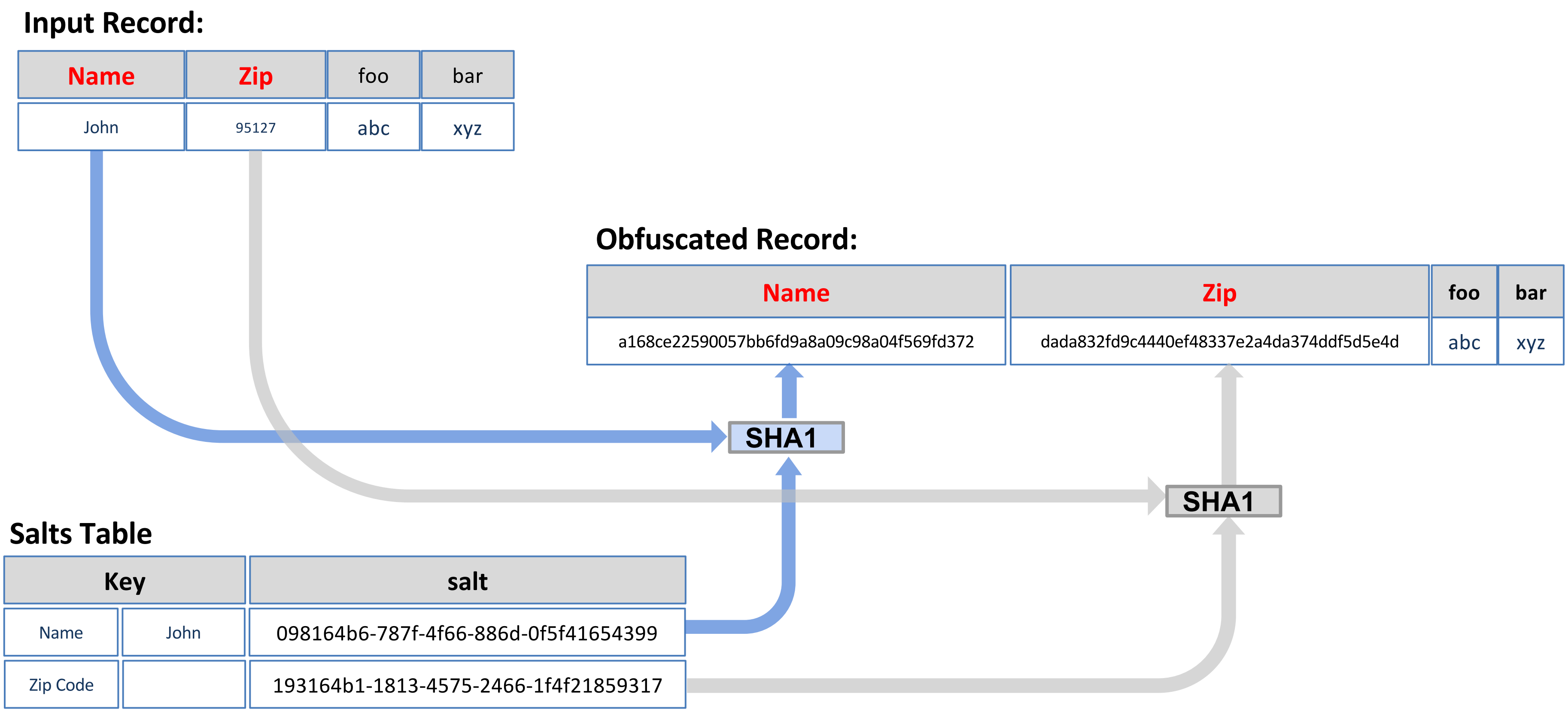
Key		salt

	Name	Zip	foo	bar
r_1	John	95127	abc	xyz

	Key		salt
[r_1]	Name	John	098164b6-787f-4f66-886d-0f5f41654399
[r_1]	Zip		193164b1-1813-4575-2466-1f4f21859317

Name	Zip	foo	bar
a168ce22590057bb6fd9a8a09c98a04f569fd372	dada832fd9c4440ef48337e2a4da374ddf5d5e4d	abc	xyz

Mapping Data Flow (cont.)



Input Record:

Name	Zip	foo	bar
John	95127	abc	xyz

Obfuscated Record:

Name	Zip	foo	bar
a168ce22590057bb6fd9a8a09c98a04f569fd372	dada832fd9c4440ef48337e2a4da374ddf5d5e4d	abc	xyz

SHA1

SHA1

Salts Table

Key		salt
Name	John	098164b6-787f-4f66-886d-0f5f41654399
Zip Code		193164b1-1813-4575-2466-1f4f21859317

http://docs.oracle.com/javase/6/docs/api/java/util/UUID.html
http://en.wikipedia.org/wiki/SHA-1

Conclusions

- Collecting, mining and analyzing large amounts of data has become a necessity in the current world

- Protecting privacy / PII is an important responsibility

- HAT provides a scalable and extensible anonymization solution to address your data privacy needs

Q&A

References

[1] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

[2] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam: l-Diversity: Privacy Beyond k-Anonymity.ICDE 2006: 24

[3] Uniqueness of Simple Demographics in the U.S. Population LIDAP-WP4 Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000 (1000) by Latanya Sweeney

[4] Privacy Technology Focus Group Report by U.S. Department of Justice - https://it.ojp.gov/documents/privacy_technology_focus_group_full_report.pdf