

# MrLazy: Lazy Runtime Label Propagation for MapReduce

**Sherif Akoush**, Lucian Carata, Ripduman Sohan, and Andy Hopper

HotCloud 2014

June 2014



# Motivation

theguardian

[News](#) | [Sport](#) | [Comment](#) | [Culture](#) | [Business](#) | [Money](#) | [Life & style](#) |

[News](#) > [Society](#) > [NHS](#)


## The NHS plan to share our medical data can save lives – but must be done right

Care.data, the grand project to make the medical records of the UK population available for scientific and commercial use, is not inherently evil – far from it – but its execution has been badly bungled. Here's how the government can regain our trust



**Ben Goldacre**

The Guardian, Friday 21 February 2014 18.30 GMT

 [Jump to comments \(311\)](#)



# England players' passport numbers revealed in teamsheet blunder

- last updated Wed 4 Jun 2014

3 TEAM OF ENGLAND				
	NO.	NAME / FIRST NAME (S) NOM DE FAMILLE / PRENOM (S) APELLIDOS / NOMBRE (S) DE PILA FAMILIENNAME / VORNAME (N)	DATE OF BIRTH DATE DE NAISSANCE FECHA DE NACIMIENTO GEBURTSDATUM	PASSPORT NO. NO DE PASSEPORT N° DE PASAPORTE PASSNUMMER
GOALKEEPER GARDIEN DE BUT GUARDAMETA TORERO	13	Foster Ben (GK)	03/04/1983	
FIELD PLAYERS JOUEURS DE CHAMP JUGADORES DE CAMPO FELDSPILER	17	Milner James	04/01/1986	
	23	Shaw Luke	12/07/1995	
	16	Jones Phil	21/02/1992	
	12	Smalling Chris	22/11/1989	
	8	Lampard Frank ©	20/06/1978	
	7	Wilshere Jack	01/01/1992	
	15	Oxlade-Chamberlain Alex	15/08/1993	
	21	Barkley Ross	05/12/1993	
	10	Rooney Wayne	24/10/1985	
	18	Lambert Rickie	16/02/1982	
RESERVES REMPLACANTS SUSTITUTOS ERSATZSPIELER	22	Forster Fraser (GK)	17/03/1988	
	14	Henderson Jordan	17/06/1990	

# Information Flow Control (IFC)

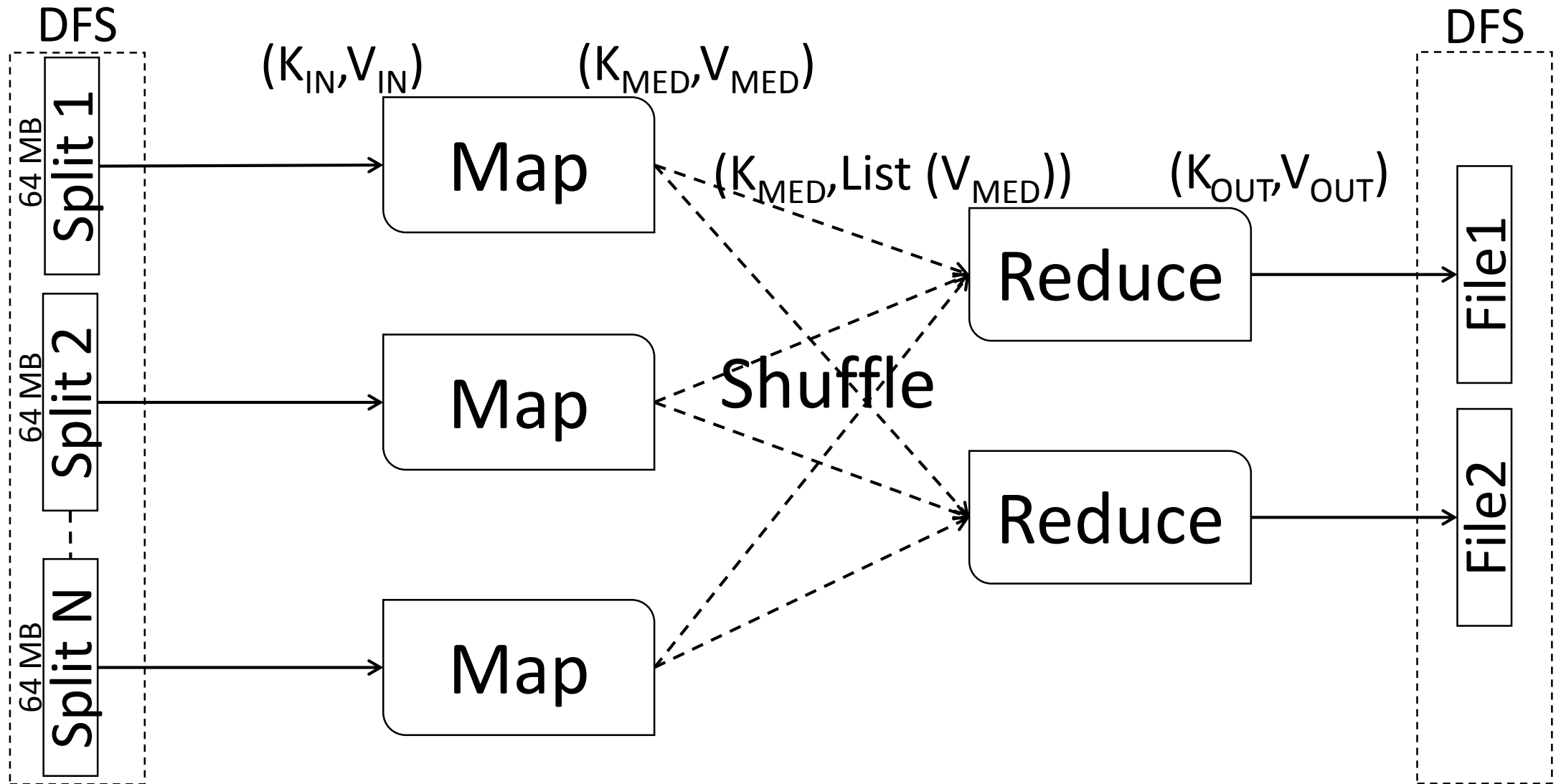
- IFC\*
  - Propagate Record + Sensitivity Metadata
  - Control Information Flow by Checking Metadata against Policies
- But...
  - Many In-House Computations
  - No Need for Active Checking
  - Only When Publishing **Some** Results
- **Lazy** IFC
  - Track and Use Lineage
  - Evaluate Output Labels When Needed

\*J. Bacon, D. Eysers, T. Pasquier, J. Singh, I. Papagiannis, and P. Pietzuch, “Information Flow Control for Secure Cloud Computing,” Network and Service Management, IEEE Transactions on, 2014.

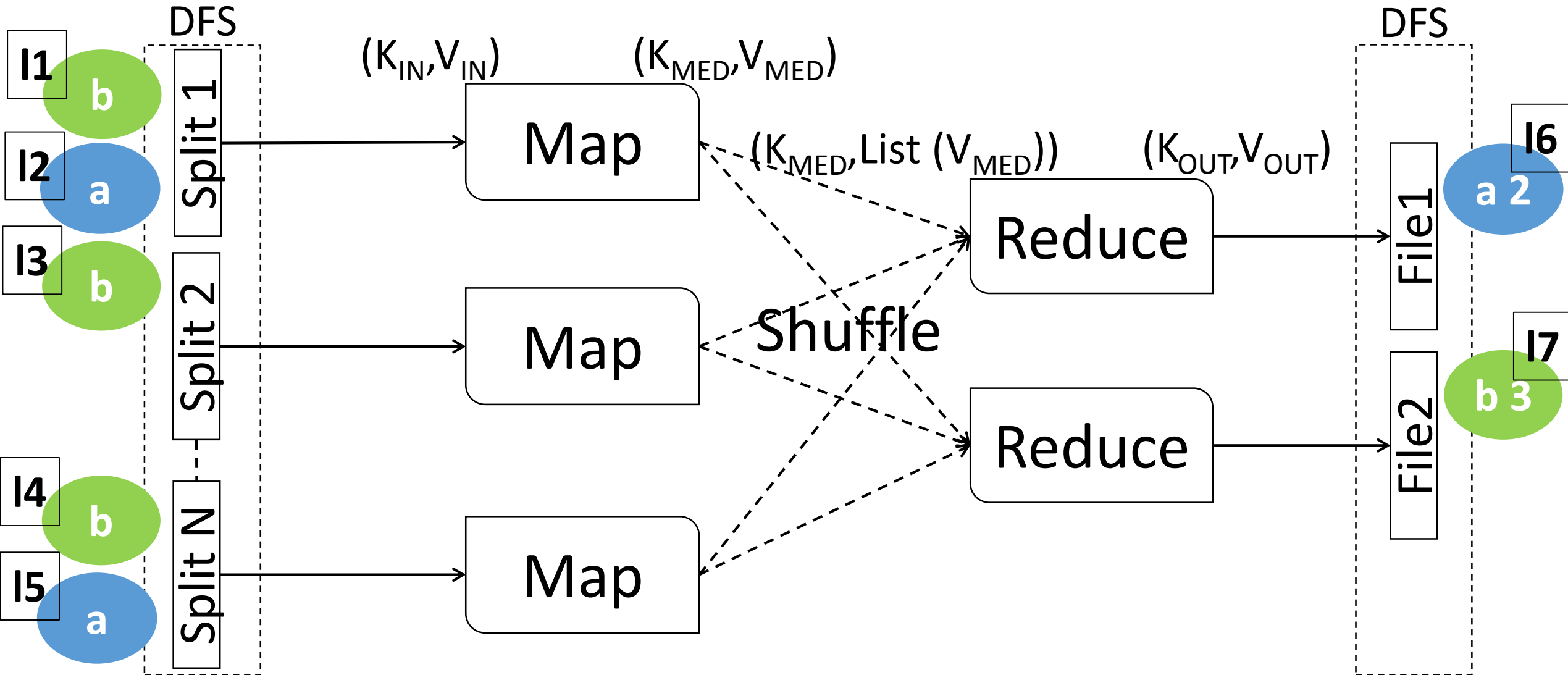
# Labels (Metadata)

- More than one Label per Record
  - Different Country Regulations, Data Quality...
- Field-Level
- Dynamic Properties
  - Users Opting In/Out
  - Sensitivity of Data Expires in 2 Years
  - New Policies

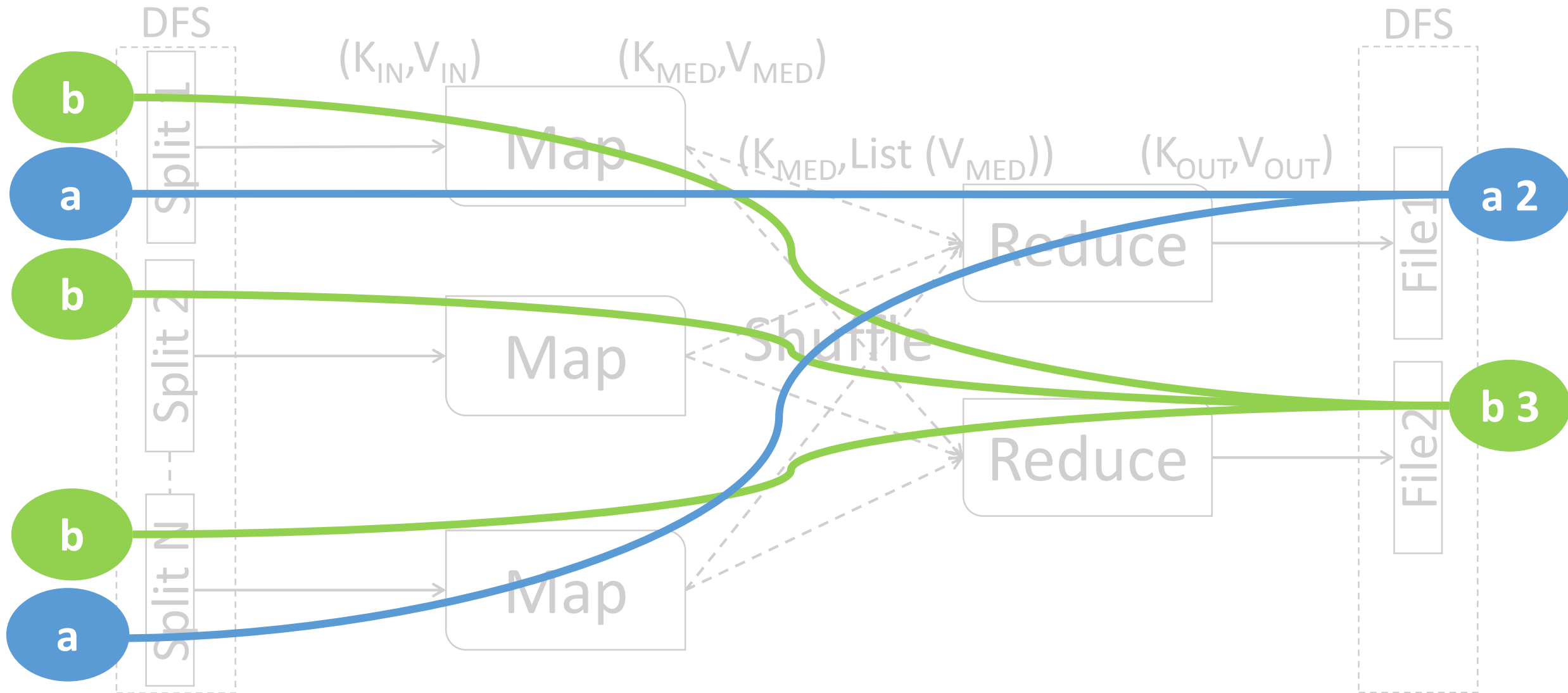
# MapReduce Paradigm



# IFC and MapReduce

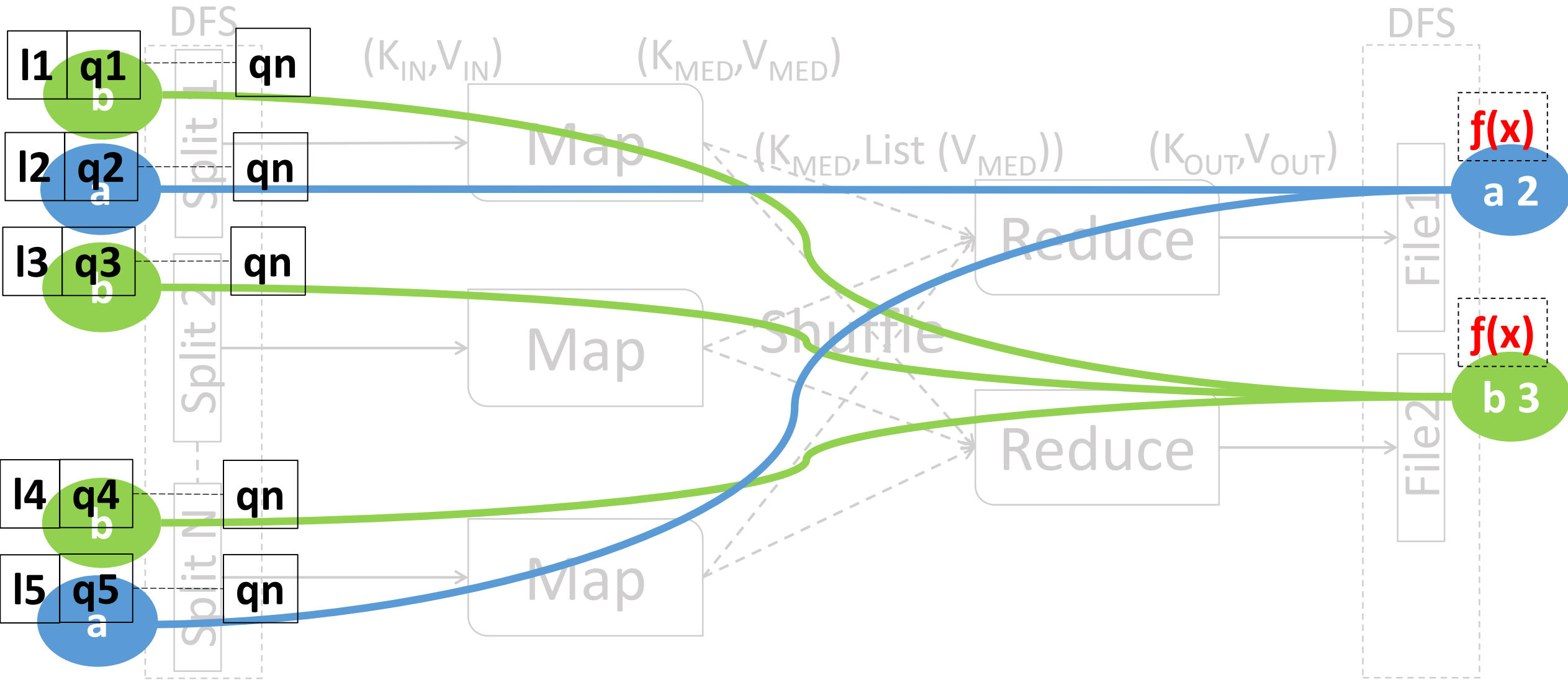


# Record-Level Lineage for MapReduce





# Lazy IFC for MapReduce



# Lineage Capture in Hadoop MapReduce

- Record-Level Lineage
- No Changes to User Code
- Always-On Feature
  - Treat Lineage for Map and Reduce Tasks **Separately**
  - Lineage Reconstruction

# Field-Level Enforcement

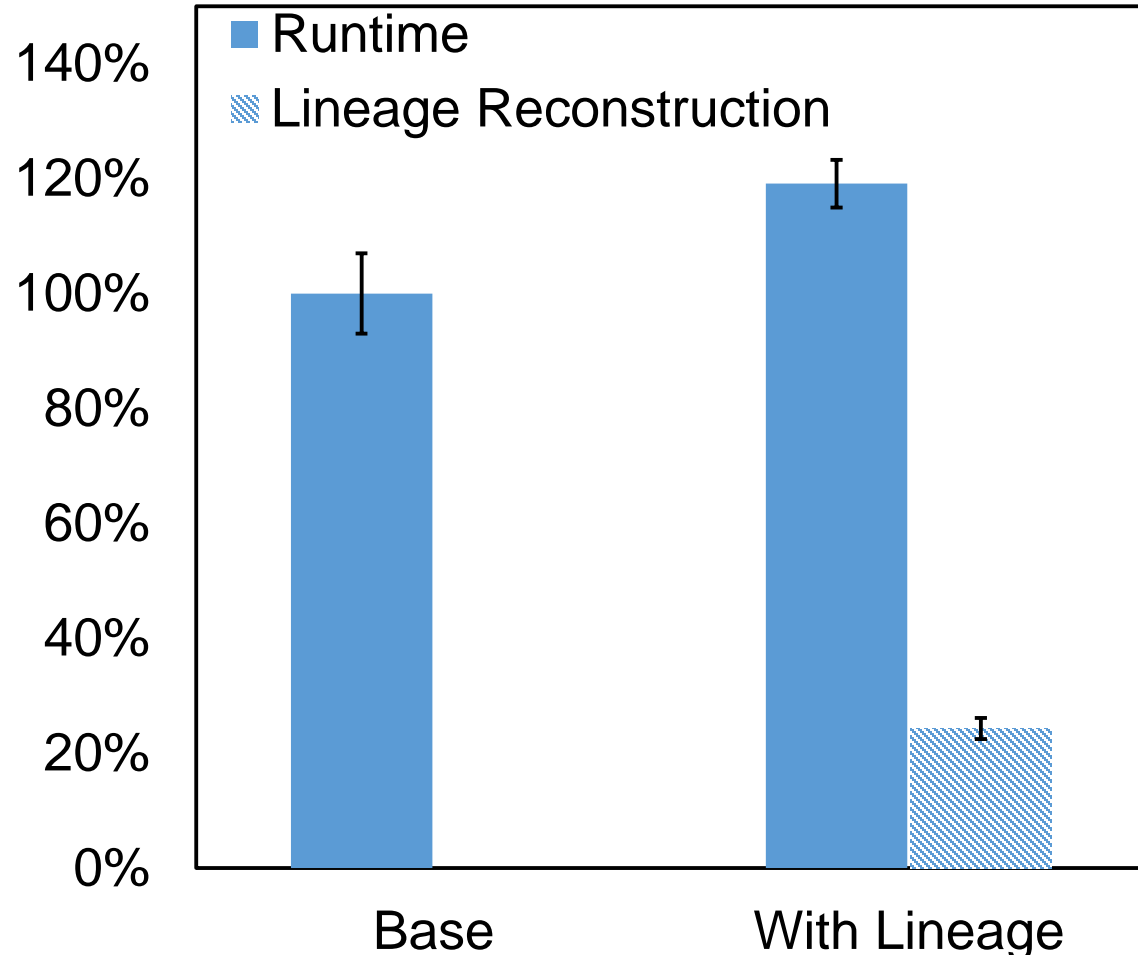
- One Record Can Have Fields With Different Sensitivity
  - Player Name vs. Passport Number
- Field-Level (Conservative) Visibility By Static Analysis

```
map(Text key, Text value)
{
    String str[] = value.toString().split(",")
    Text name = new Text(str[0])
    write(name, 1)
}
```

# Prototype Evaluation

- Implementation in Hadoop MapReduce
- 7-node Cluster
- Dataset from BigDataBench: 120 GB
- Join and Filter Job

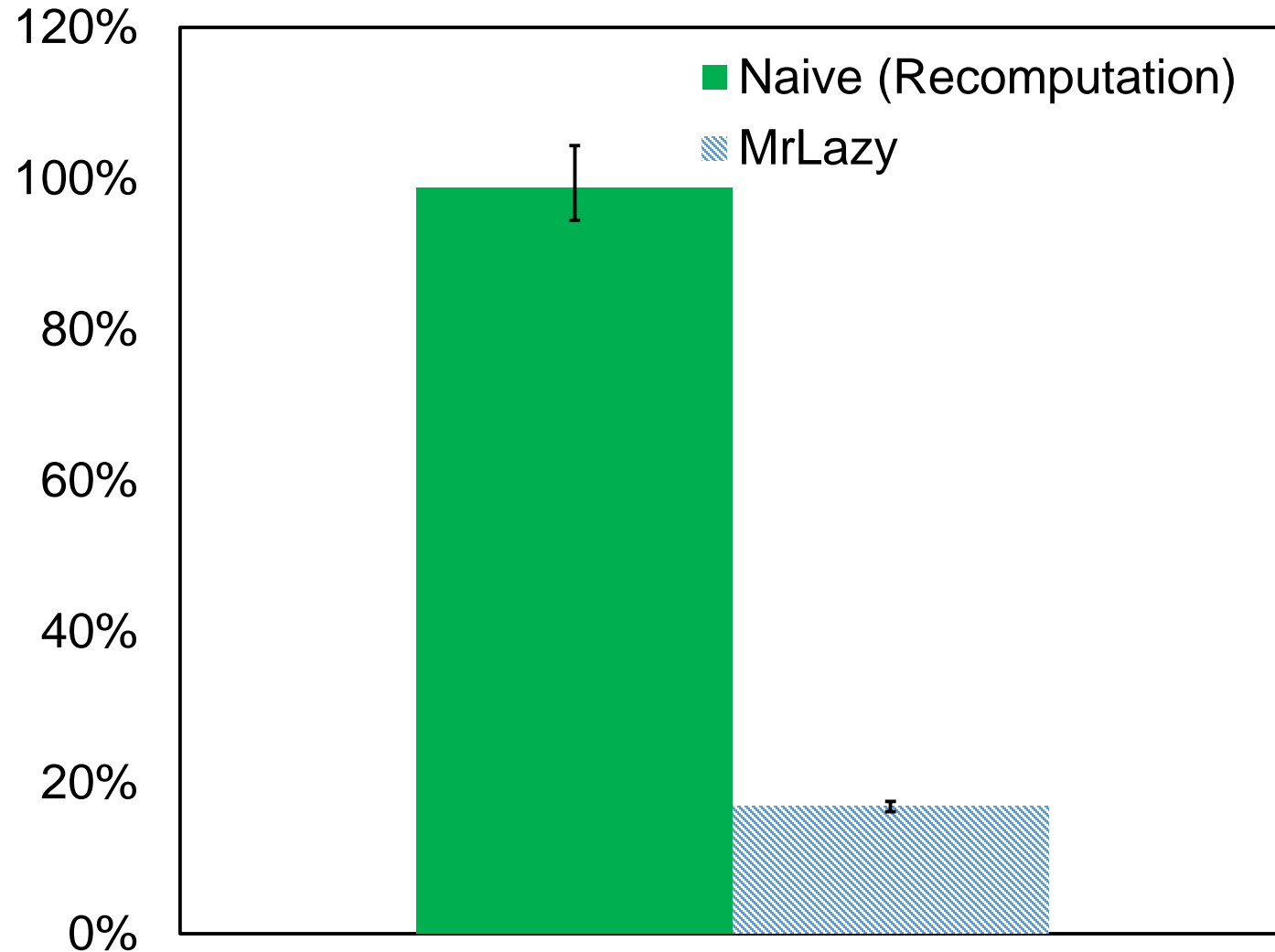
# Overheads (Lineage Capture)



- Storage
  - 50% of Output
  - Delete When Not Needed
  - Trading Space for Time

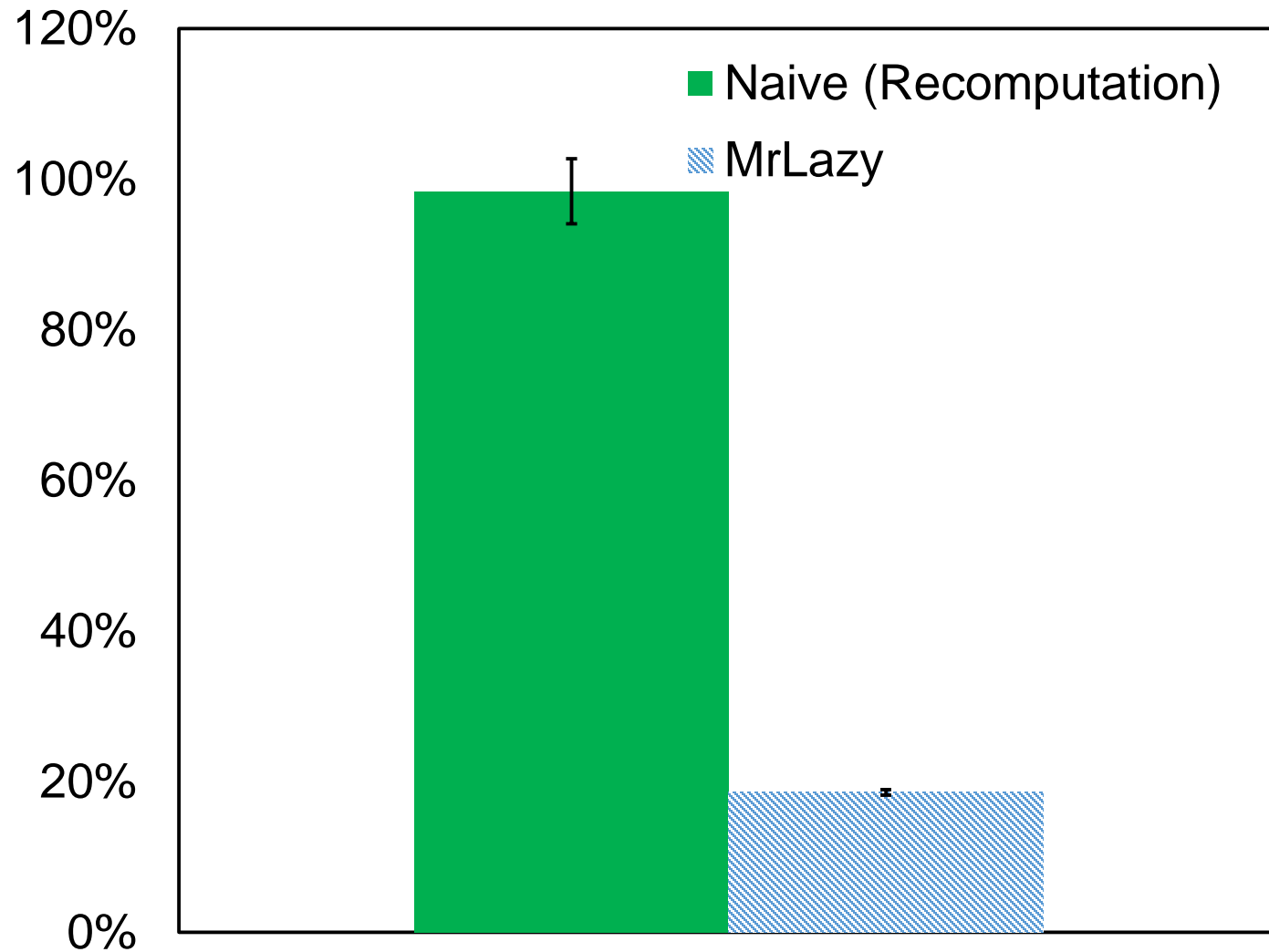
# Policy 1: Users Opt-out of Data Sharing

- 5% of Users



# Policy 2: Sensitivity of Data Lasts 2 Years

- Dynamic Behaviour



# Other Challenges

- Dealing with State
  - In-lining Instructions to Expose State
  - TopK
- Subtle Data Leakage
  - Differential Privacy



# Conclusion

- Delay Output Label (Metadata) Computation
- Fine-Grained Lineage as Audit Mechanism
- Non-Prohibitive Overheads
- Future Work:
  - Reducing Overheads
  - Large-Scale Evaluation
  - Recomputation-Based Recovery from Failures

# Thanks

[Sherif.Akoush@cl.cam.ac.uk](mailto:Sherif.Akoush@cl.cam.ac.uk)

<http://www.cl.cam.ac.uk/~sa497/>