CLaDS:

A Cloud-Based Virtual Lab for the Delivery of Scalable Hands-on Assignments for Practical Data Science Education

Chase Geigle, Ismini Lourentzou, Hari Sundaram, and ChengXiang Zhai

2018-07-03

University of Illinois at Urbana-Champaign

Demand for Data Science/Analytics

2,350,000 job listings in 2015

Demand for **data scientists and data engineers** is expected to grow by **39%** by 2020

Skill Name	Predicted 2-Year Growth
Data Science	93%
Machine Learning	56%
Tableau	52%
Big Data	50%
Data Visualization	44%

`https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IML14576USEN`

Academia's Response: Online Masters?

A few examples (non-exhaustive):

- UC Berkeley: **Master of Information and Data Science** (2014)
- UIUC: **Master of Computer Science in Data Science** (Coursera, 2017)
- UMich: **Master of Applied Data Science** (Coursera, coming 2019)
- "a top-tier university": **Master of Data Science** (edX, coming "soon")

Problem: how do we scale an online masters program to a thousand students?

What should **providing computing resources** look like?

Challenges of High-Scalability Data Science Education

Assignments to train the next generation of data scientists must:

1. scale to **large numbers of students**,
2. be able to cover a **broad spectrum** of potential skills,
3. allow for the development of **hands-on experience** with **real data sets**, and
4. **minimize** overall deployment **cost**.

CLaDS: A Cloud-based Lab for Data Science

Related Work

Auto-gradable programming assignments:

- Mike Joy, Nathan Griffiths, and Russell Boyatt. “The boss online submission and assessment system”. In: *Journal on Educational Resources in Computing (JERIC)* 5.3 (2005), p. 2
- Christoph Matthies, Arian Treffer, and Matthias Uflacker. “Prof. CI: Employing continuous integration services and Github workflows to teach test-driven development”. In: *2017 IEEE Frontiers in Education Conference (FIE)*. IEEE. 2017, pp. 1–8
- `www.hackerrank.com`, `www.topcoder.com`

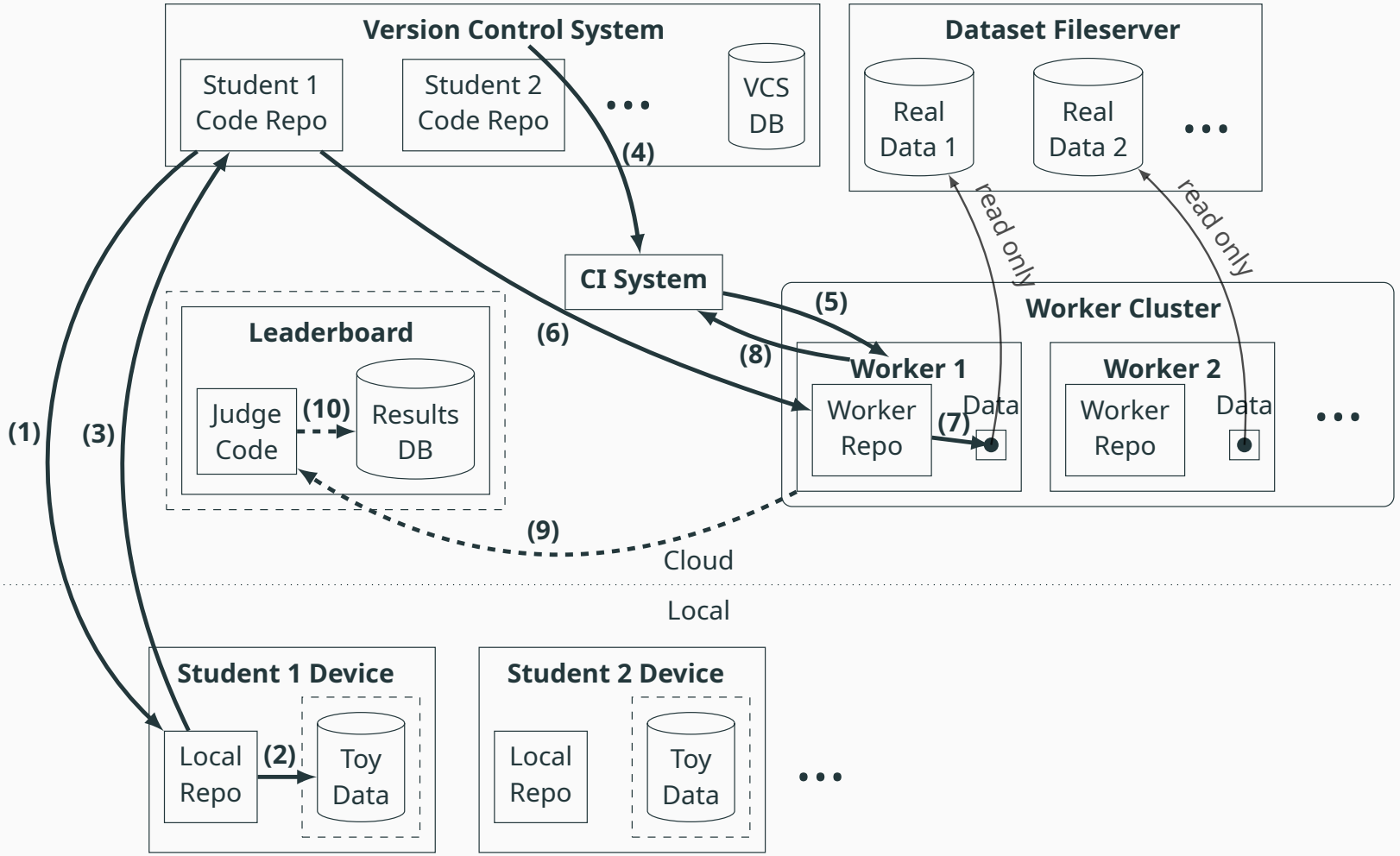
Closest idea: Kaggle (`www.kaggle.com`); key differences:

- complete flexibility (tools, libraries, grading, competition vs. traditional assignments)
- no hard limit on the size of dataset

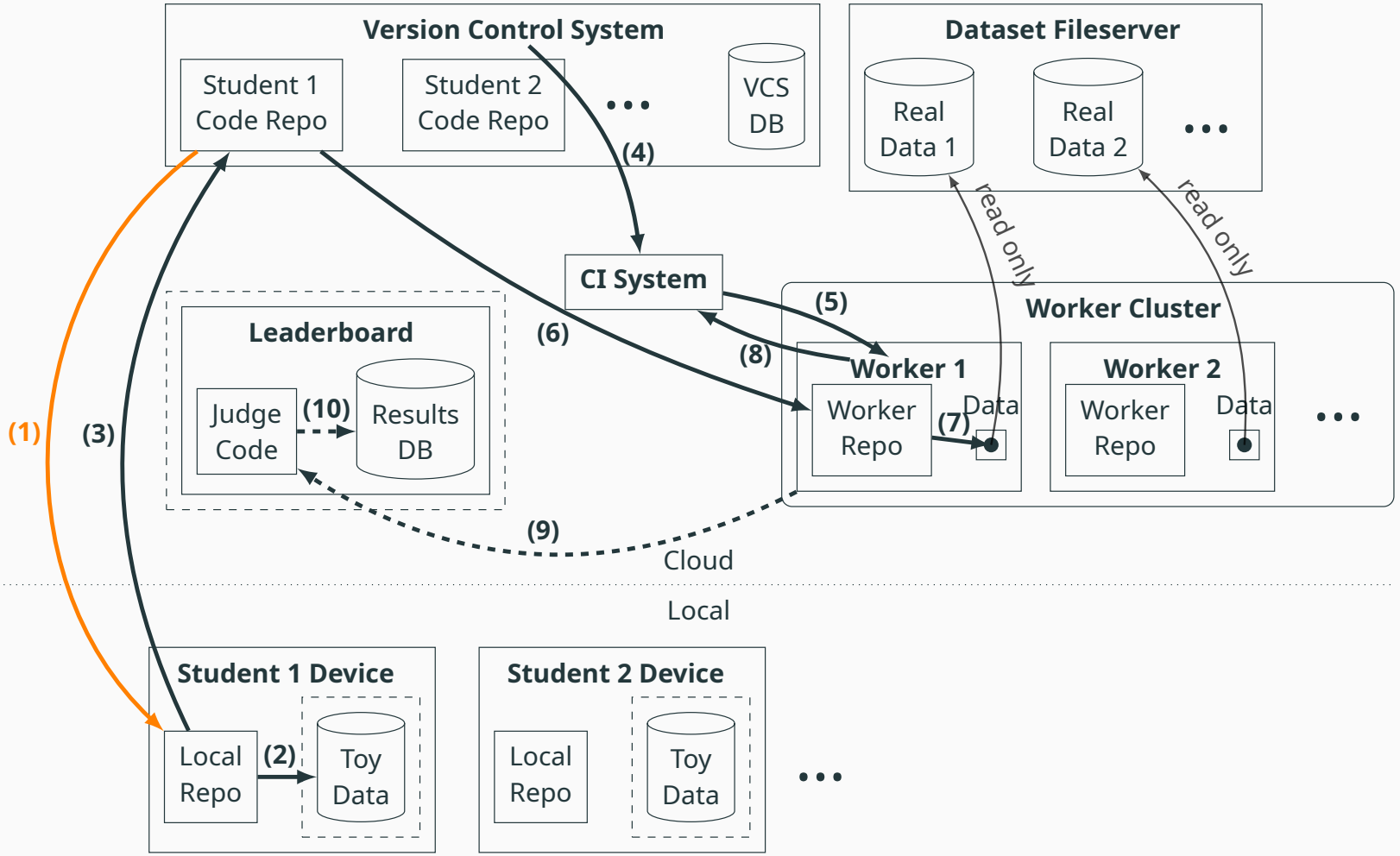
Domain-specific labs (CLaDS is general):

- Hui Fang et al. “VIRLab: A Web-based Virtual Lab for Learning and Studying Information Retrieval Models”. In: *Proc. SIGIR*. Gold Coast, Queensland, Australia: ACM, 2014, pp. 1249–1250
- Hui Fang and ChengXiang Zhai. “VIRLab: A Platform for Privacy-Preserving Evaluation for Information Retrieval Models”. In: *Proc. PIR@SIGIR*. Gold Cost, Queensland, Australia, 2014, pp. 37–38

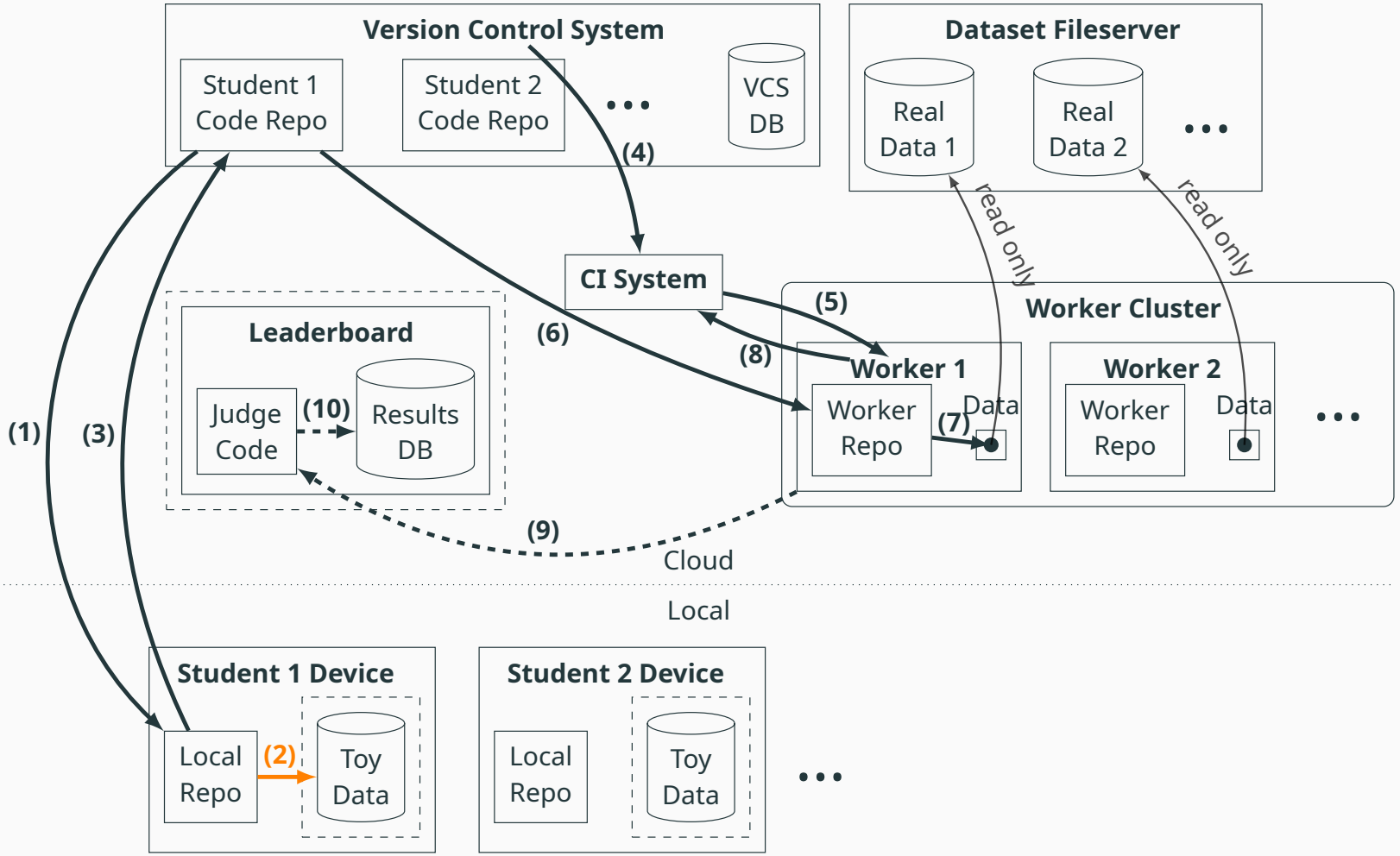
System Architecture



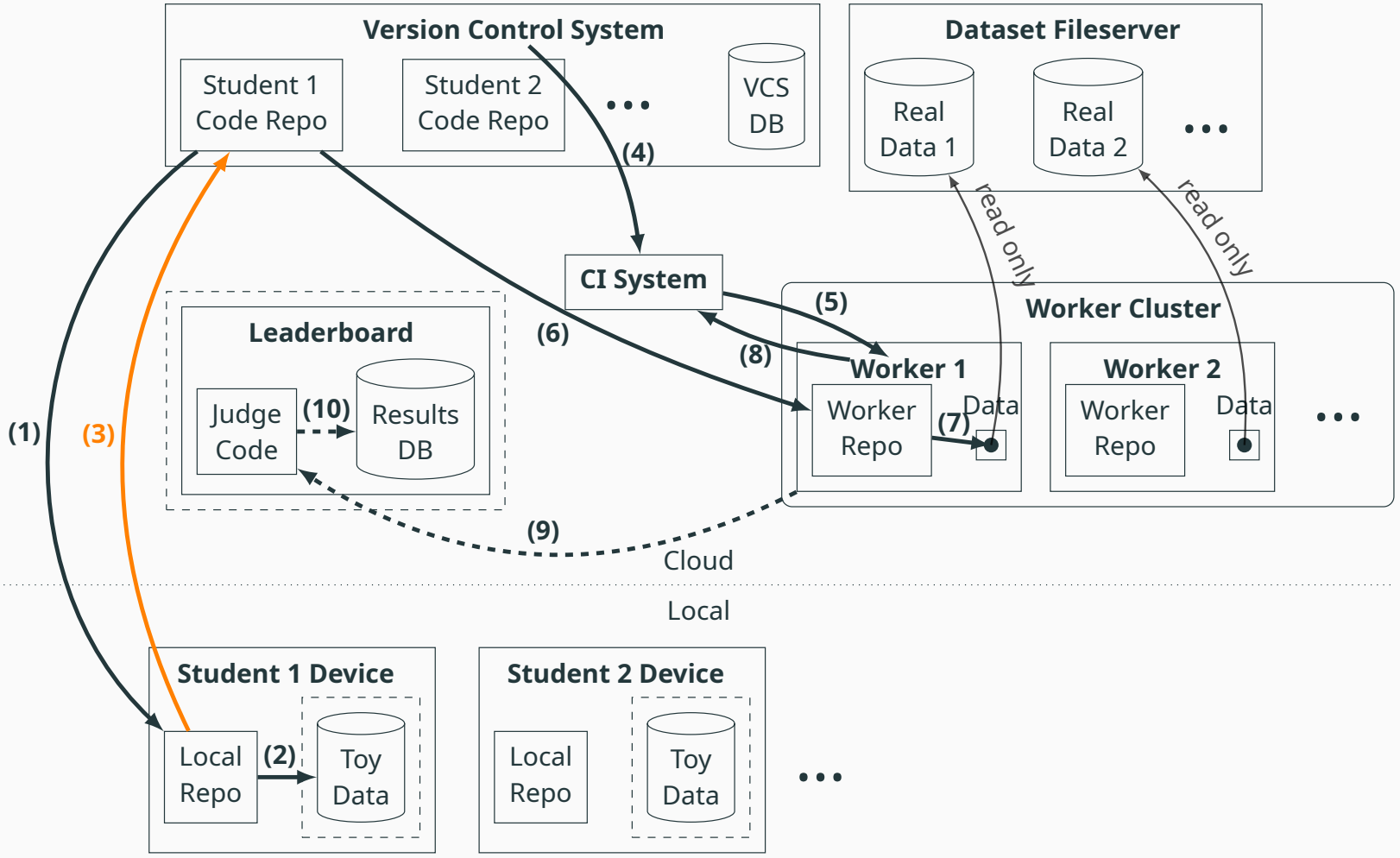
System Architecture



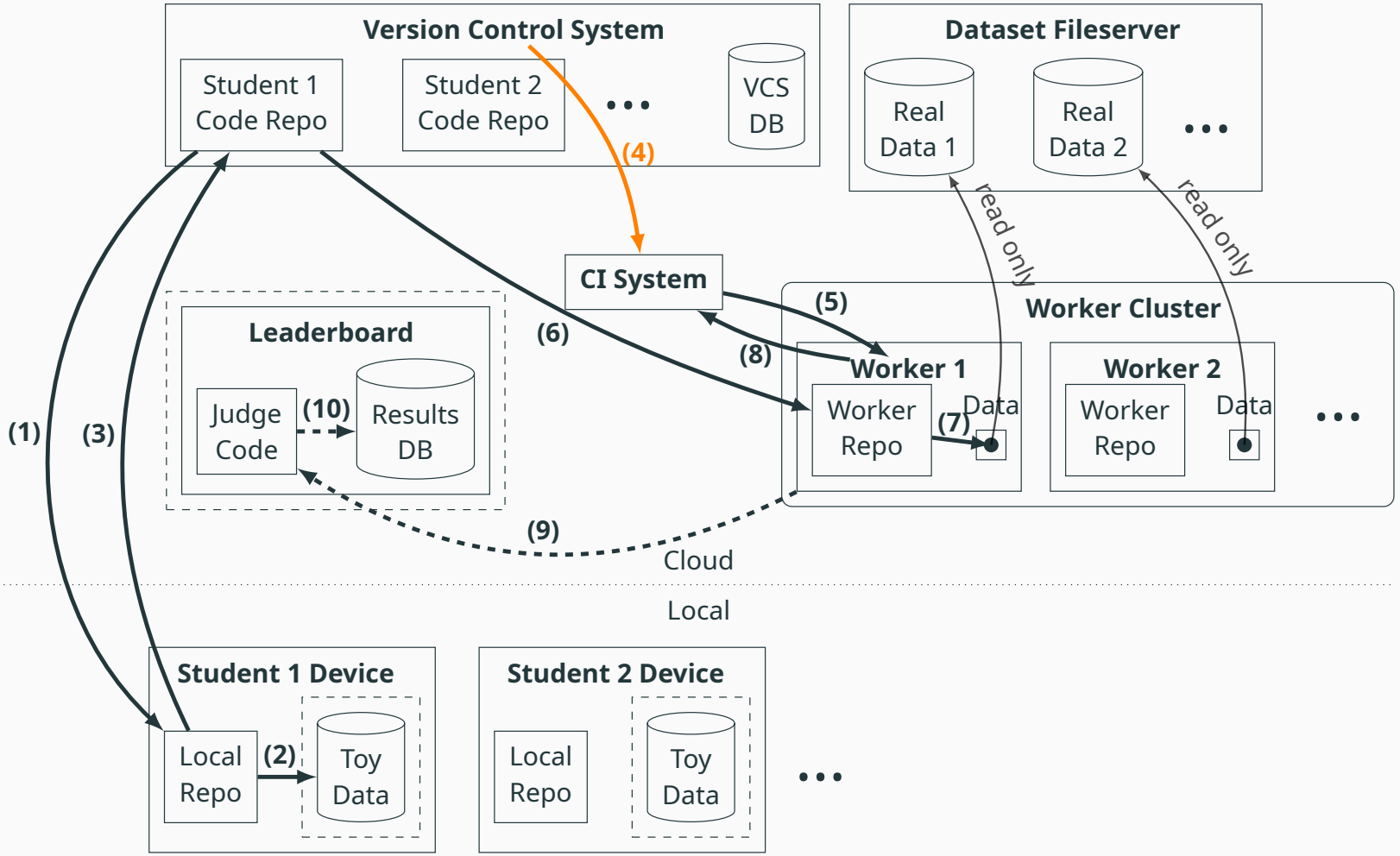
System Architecture



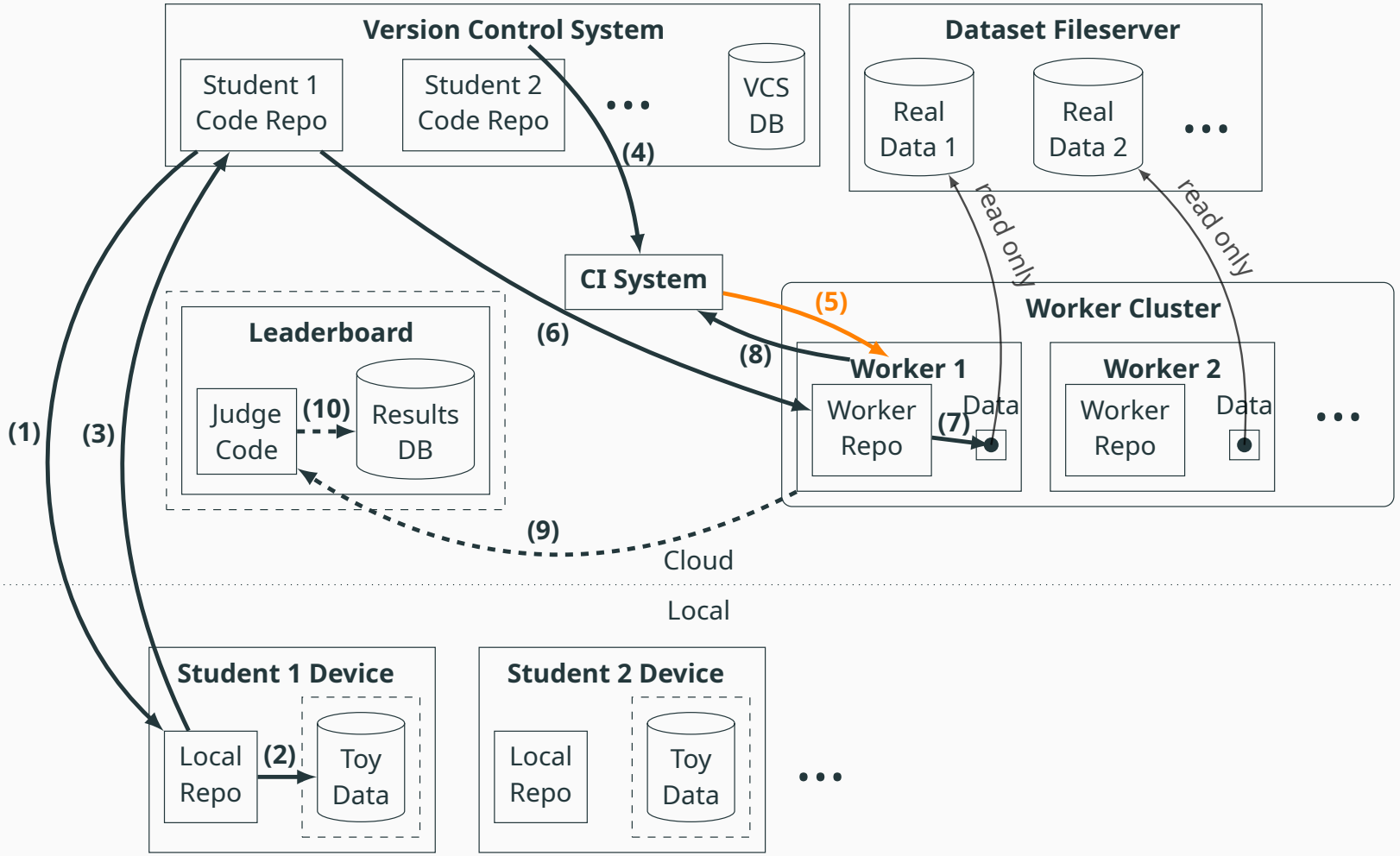
System Architecture



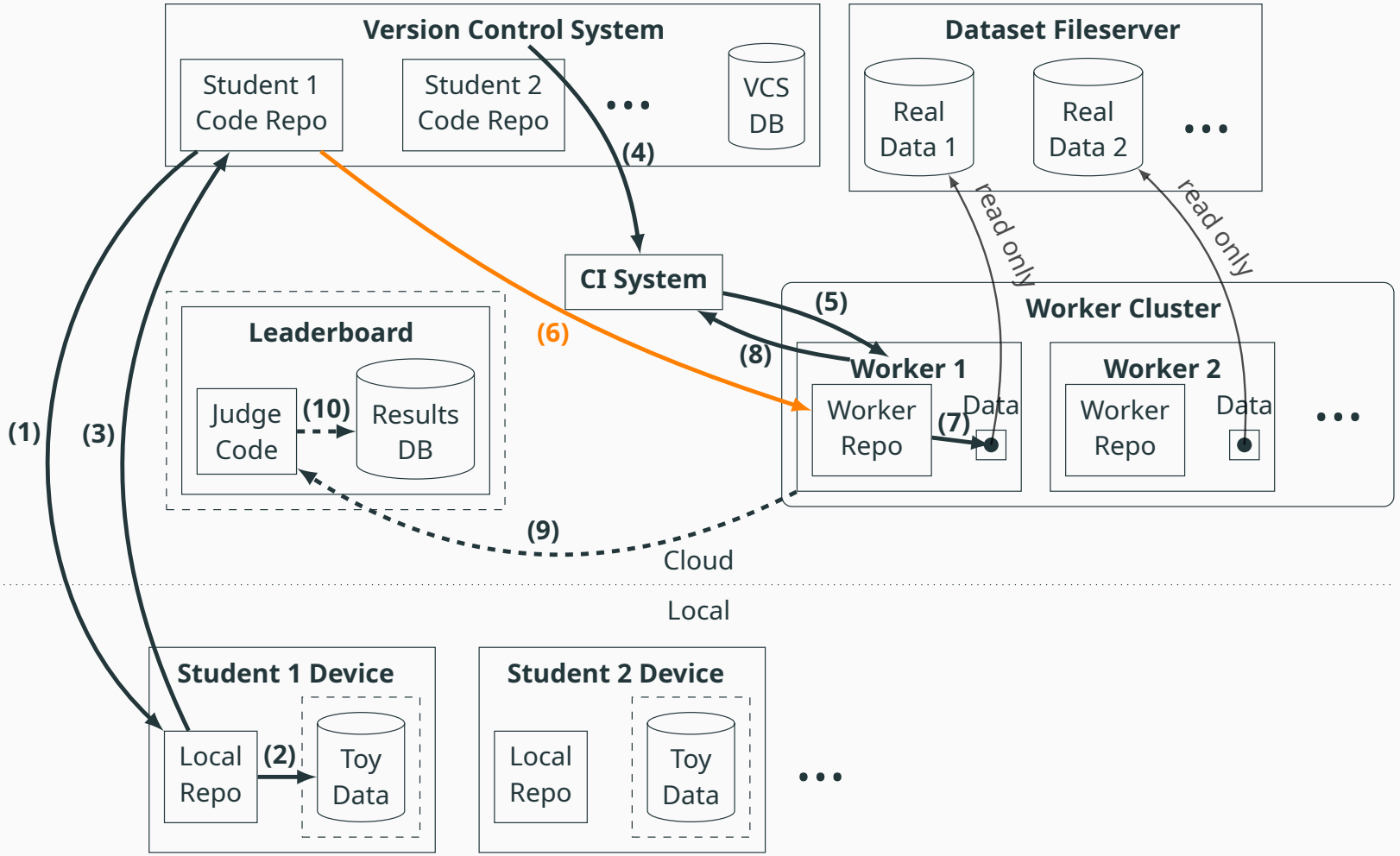
System Architecture



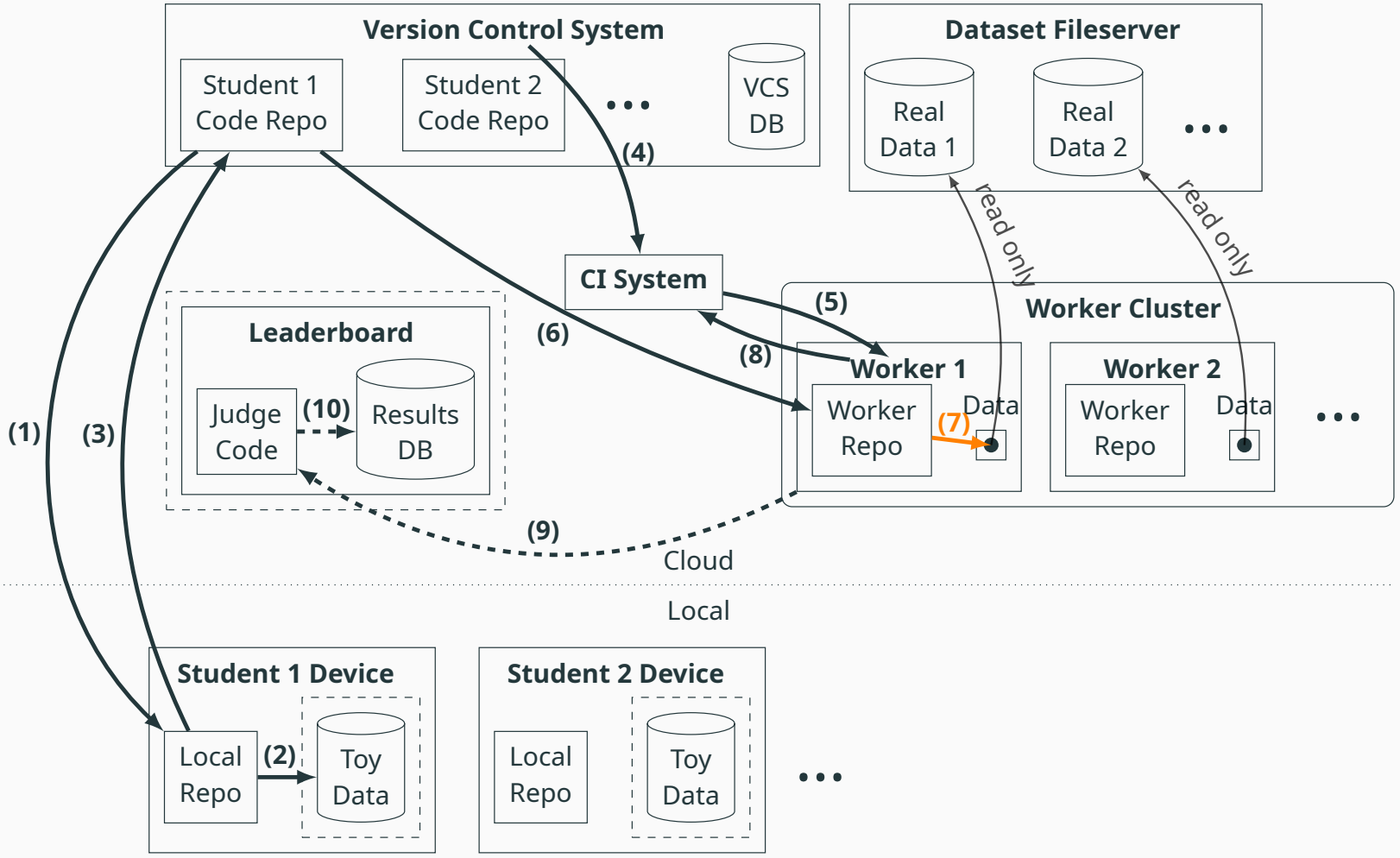
System Architecture



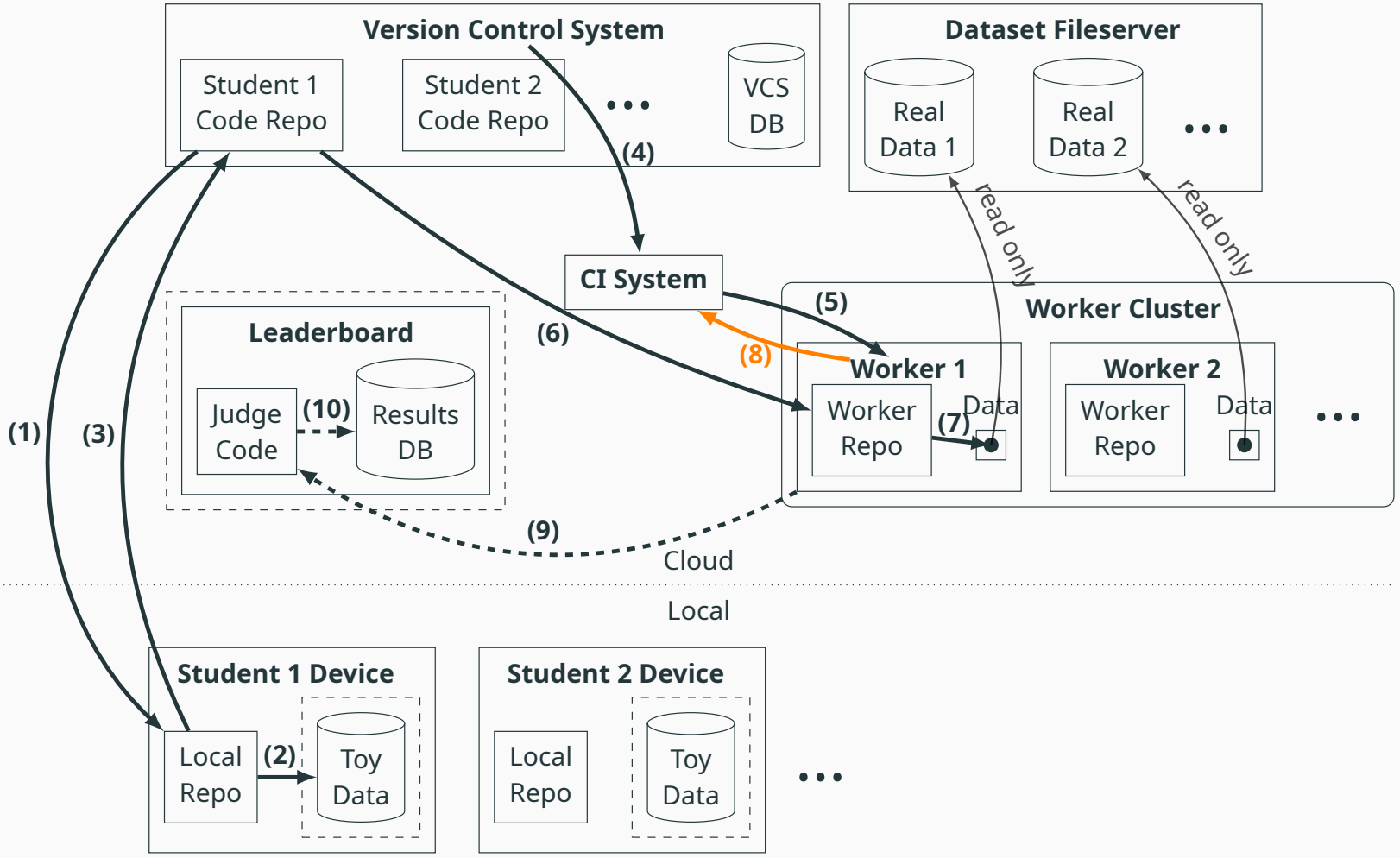
System Architecture



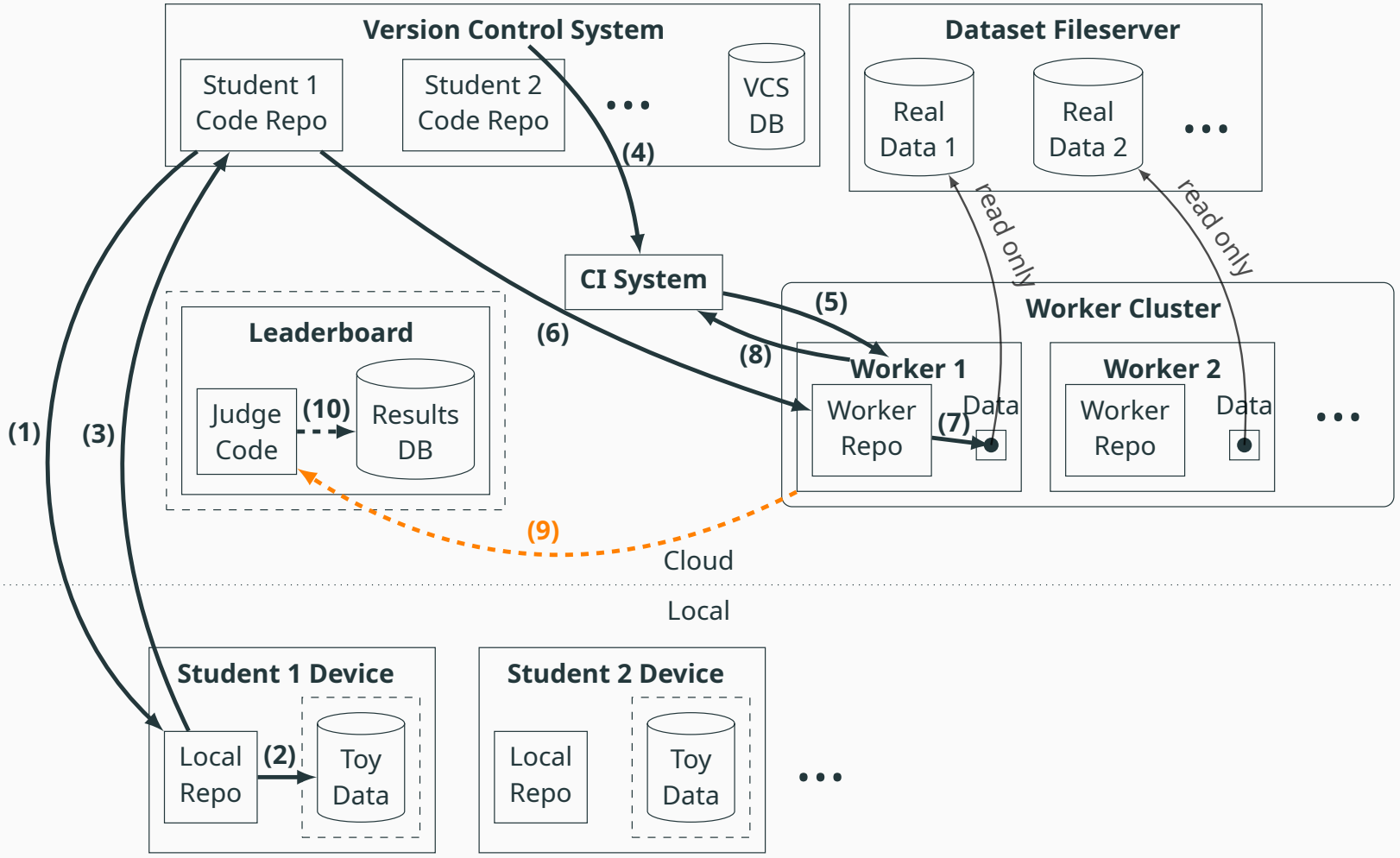
System Architecture



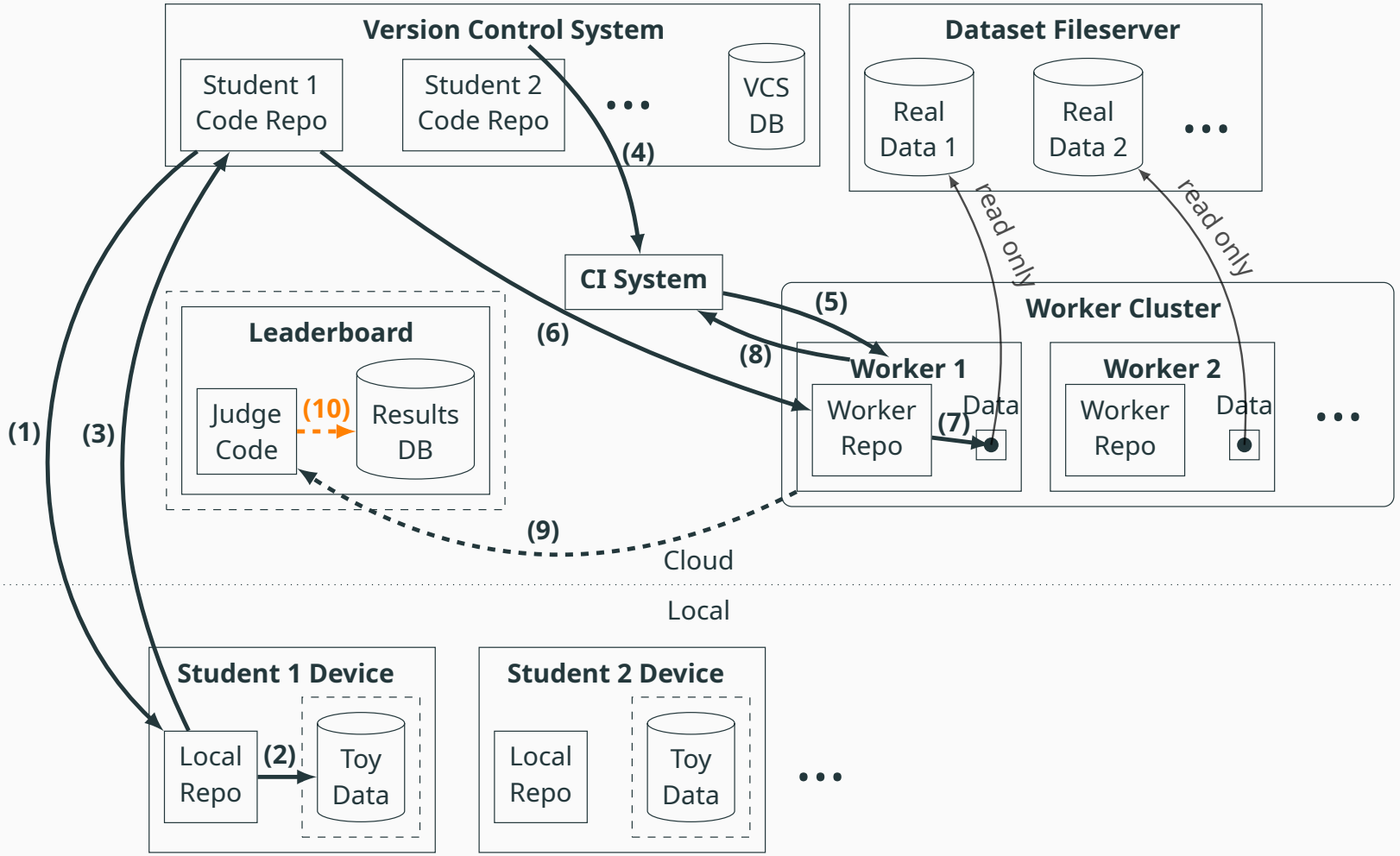
System Architecture



System Architecture



System Architecture



Key Components

What makes it work?

- Auto-scaling worker cluster: computing resources are **demand driven**
- Assignment agnostic: supports a wide variety of **assignment types**
- Workers live “next to” datasets: opens door to **real**(er) **datasets**
- Simpler, event-driven (optional) leaderboards

Software:

- GitLab (VCS; open-core)
- GitLab CI (CI system; open-core)
- Leaderboards (open-source)
- `https://timan-group.github.io/clads/`

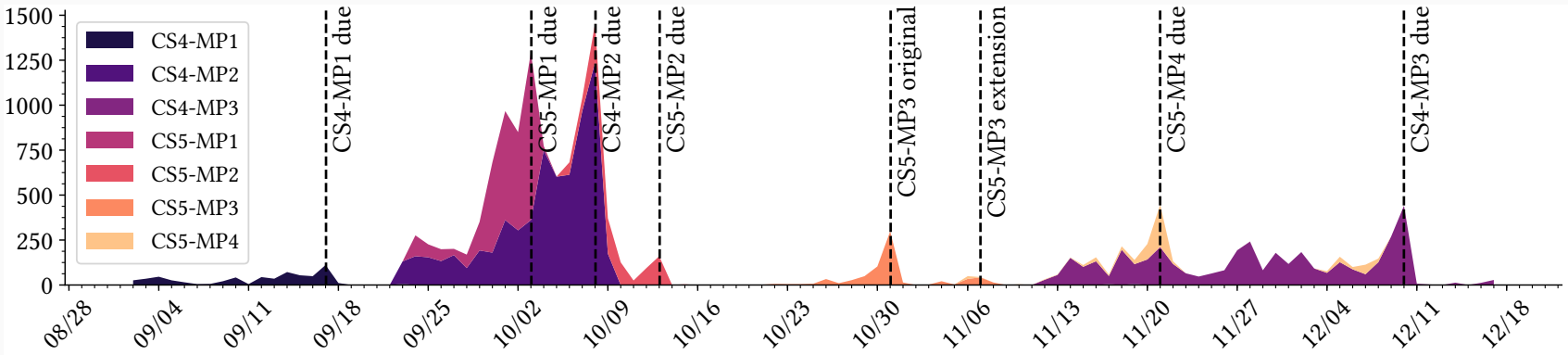
Deployment Experience

Fall 2017:

- CS 410 (online, 136 students)
- CS 510 (on-campus, 91 students)
- **Amortized cost per student: $7.40 USD**

System Utilization

	CS4		CS5	
	Description	Hrs.	Description	Hrs.
MP1	feat. extraction	28.6	smoothing methods	908.8
MP2	retrieval fns.	227.1	word embeddings	79.3
MP3	classification	142.9	topic mdls.	250.9
MP4	-	-	hidden Markov mdls.	29.6



Search Engine Competition: Leaderboard

Search Engine Competition

Rank	Alias	Overall Score		apnews (0.7)		cranfield (0.3)		Updated	Submissions
		Current	Previous	NDCG@10	Previous	NDCG@10	Previous		
1	我跟你讲啊，你们这样子在测试集上调参啊，是不行的！	0.7967	*0.7967*	0.99	*0.99*	0.3455	*0.3455*	2018-02-25 | 19:44:53	18
2	Alpaca_cc	0.4166	*0.4166*	0.438	*0.438*	0.3666	*0.3666*	2018-02-21 | 14:06:00	5
3	gh	0.4134	*0.4209*	0.4338	*0.446*	0.3657	*0.3625*	2018-02-20 | 22:37:32	46
3	Rohan	0.4134	*0.4134*	0.4338	*0.4338*	0.3657	*0.3657*	2018-02-20 | 22:57:41	44
5	oneDayMore	0.3987	*0.3987*	0.4199	*0.4199*	0.3493	*0.3493*	2018-02-21 | 13:26:30	20
6	xingye	0.3986	*0.3986*	0.4207	*0.4207*	0.3471	*0.3471*	2018-02-18 | 21:49:10	34
7	bwj	0.3985	*0.3985*	0.4243	*0.4243*	0.3383	*0.3383*	2018-02-18 | 23:03:30	182
7	BBH	0.3985	*0.3984*	0.4243	*0.4243*	0.3383	*0.3382*	2018-02-18 | 23:48:45	81
9	hrukalive	0.3984	*0.3984*	0.4194	*0.4194*	0.3495	*0.3495*	2018-02-21 | 21:22:01	33
10	Joey	0.3984	*0.3984*	0.4243	*0.4241*	0.3381	*0.3382*	2018-02-18 | 23:04:51	104
11	lily	0.3983	*0.3983*	0.4209	*0.4209*	0.3457	*0.3457*	2018-02-21 | 14:29:47	11
12	ThomasMuller25	0.3979	*0.3978*	0.4177	*0.4177*	0.3516	*0.3515*	2018-02-25 | 23:59:25	96
13	BillyisntBilibili	0.3972	*0.3992*	0.414	*0.4207*	0.358	*0.3492*	2018-02-22 | 21:54:01	32
14	Try one try	0.397	*0.3884*	0.4245	*0.4046*	0.3328	*0.3508*	2018-02-21 | 23:28:07	76

197	cshang	0.3765	*0.3668*	0.3879	*0.3773*	0.35	*0.3424*	2018-02-18 | 22:18:40	4
198	MoNeY_Pro	0.3754	*-inf*	0.386	*-inf*	0.3506	*-inf*	2018-02-16 | 15:50:36	1
198	sliu134	0.3754	*0.3754*	0.386	*0.386*	0.3506	*0.3506*	2018-02-18 | 17:42:27	3
200	Baseline	0.3702	*-inf*	0.3806	*-inf*	0.3461	*-inf*	2017-03-02 | 01:03:25	1
200	Chongye	0.3702	*0.3702*	0.3806	*0.3806*	0.3461	*0.3461*	2018-02-15 | 15:21:31	9
200	Anonymous	0.3702	*0.3702*	0.3806	*0.3806*	0.3461	*0.3461*	2018-02-18 | 23:59:56	3
200	Anonymous	0.3702	*-inf*	0.3806	*-inf*	0.3461	*-inf*	2018-02-18 | 18:43:50	1

Competition Submissions *After Beating Baseline*

Assignment: beat a baseline solution for “A”; compete for extra credit (top X positions in leaderboard).

Table shows **additional effort above and beyond** passing the assignment with a perfect score, even in the 25th percentile.

Assignment	Mean	Std. Dev.	Median	25th %ile
CS4-MP2	20.5	27.6	10.0	5.0
CS4-MP3	21.7	42.3	10.0	3.0

Student Thoughts

- “Professionally speaking, I really feel that I gained a lot, as now I truly understand the essential fundamentals in Text Information Systems areas and, thanks to the hands-on final project, and MPs, can implement some of these principles. I am more than positive that I will utilize the gained knowledge in my workplace in 2018 and make a significant impact.”

- “The system used for the programming assignments to automatically test, evaluate, and rank solutions made the assignments a fun challenge.”

- “The competition style leader board added a fun aspect…”

Takeaways

1. Now possible to perform hands-on data analysis experiments **at scale**

2. **Reproducible research**: archive of methods and their performance alongside public leaderboards for standard tasks

3. Applicable to **more than just classroom learning**—industry training, internal research groups

4. Future of computing resource provisioning for (non-sensitive) engineering work? **Use the cloud for what it's good at!**

Limitations and Future Work

Everything has a limit: just how big can the datasets get?

Everything has a limit: just how many students can we service? (Try MOOC-scale?)

Branch out: ought to be usable by other CS domains—not only data science

Security: how to prevent abuse? Must be addressed for MOOC-scale

Funding: will industry always support this? lab fees? nonprofit org?

Thank you!