Global Analytics in the Face of Bandwidth and Regulatory Constraints

Vulimiri et al.

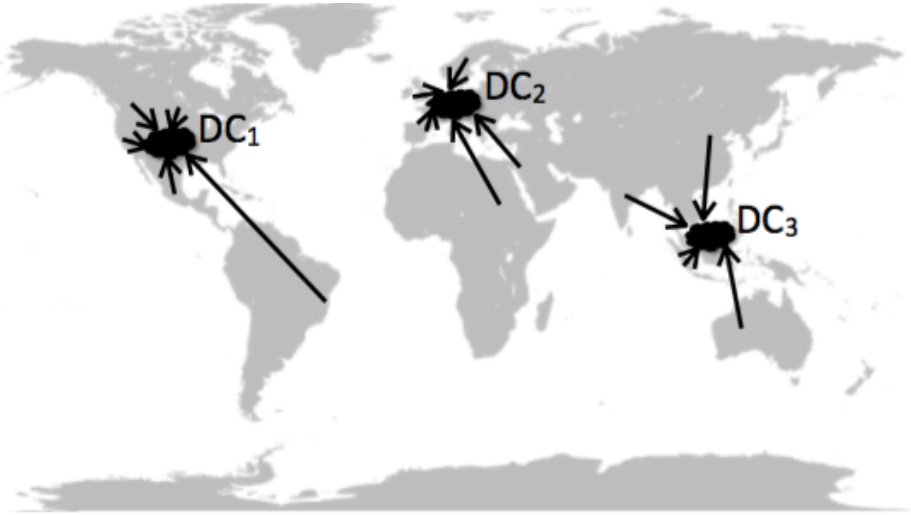
NSDI'15

Massive Data Volume

Facebook	600 TB/day
Twitter	100 TB/day
Microsoft	10s TB/day
LinkedIn	10 TB/day
Yahoo!	10 TB/day

Use case:

- User activity logs
- Monitoring remote infrastructure
- …



Collected across several data centers for low user latency

SQL analytics across geo-distributed data to extract insight

Consider an analytical query reporting statistics of users generating at least $100 in ad revenue

SQL analytics across geo-distributed data to extract insight

ClickLog
(source IP,
destURL,
visitData,
adRevenue,
…)

SQL analytics across geo-distributed data to extract insight

PageInfo
(pageURL,
pageSize,
pageRank, …)

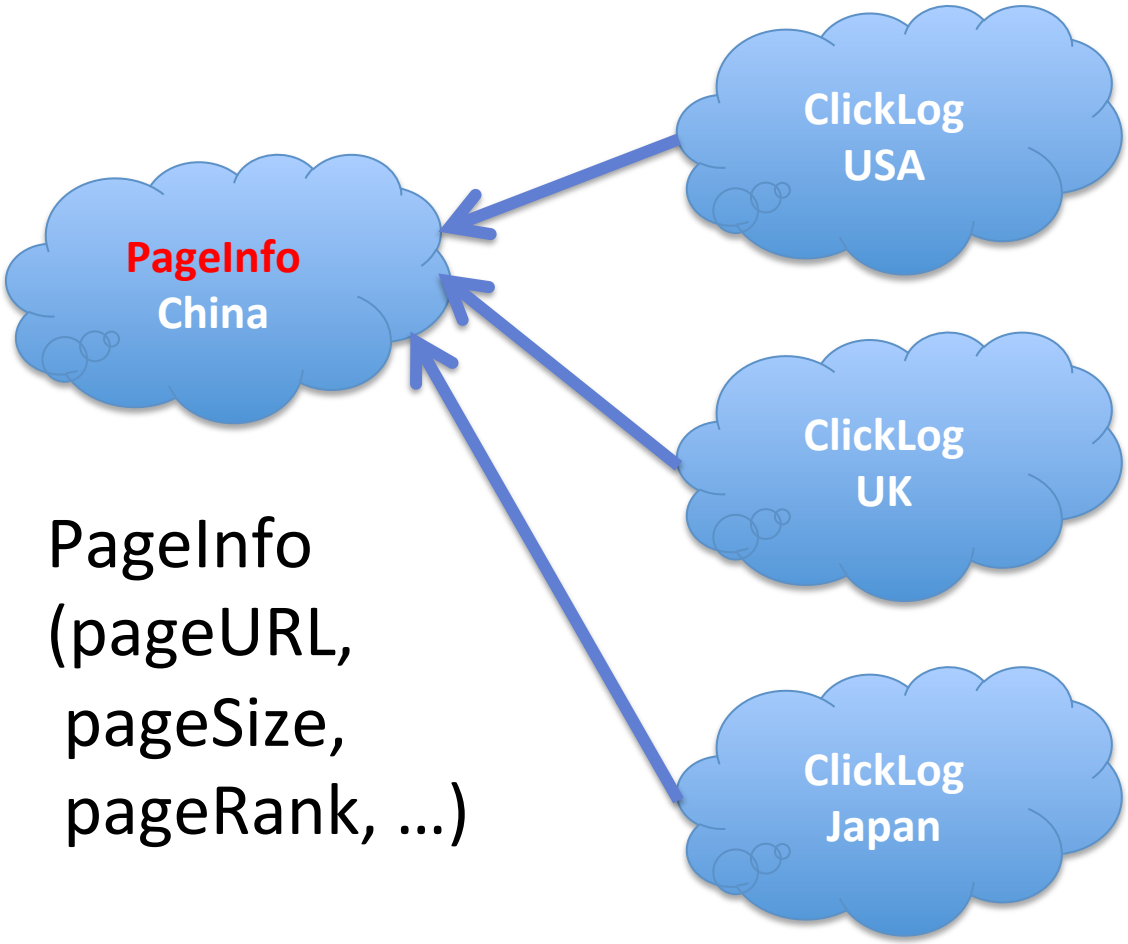
ClickLog
(source IP,
destURL,
visitData,
adRevenue,
…)

SQL analytics across geo-distributed data to extract insight

ClickLog (source IP, destURL, visitData, adRevenue, …)
PageInfo (pageURL, pageSize, pageRank, …)

Q: SELECT sourceIP, sum(adRevenue), avg(pageRank)
FROM ClickLog cl JOIN PageInfo pi
ON cl.destURL = pi.pageURL
WHERE pi.pageCategory = 'Entertainment'
GROUP BY sourceIP
HAVING sum(adRevenue) >= 100

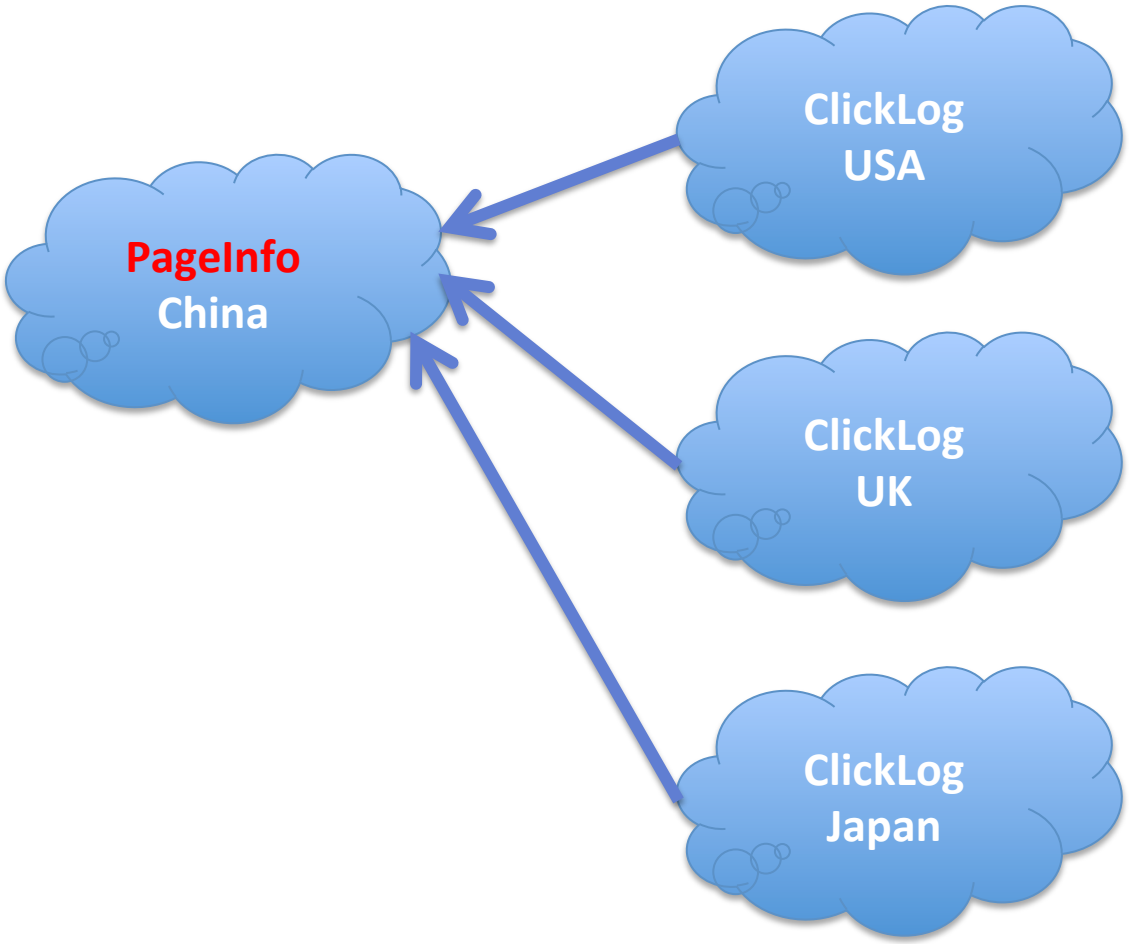
Centralized method



PageInfo
(pageURL,
pageSize,
pageRank, …)

ClickLog
(source IP,
destURL,
visitData,
adRevenue,
…)

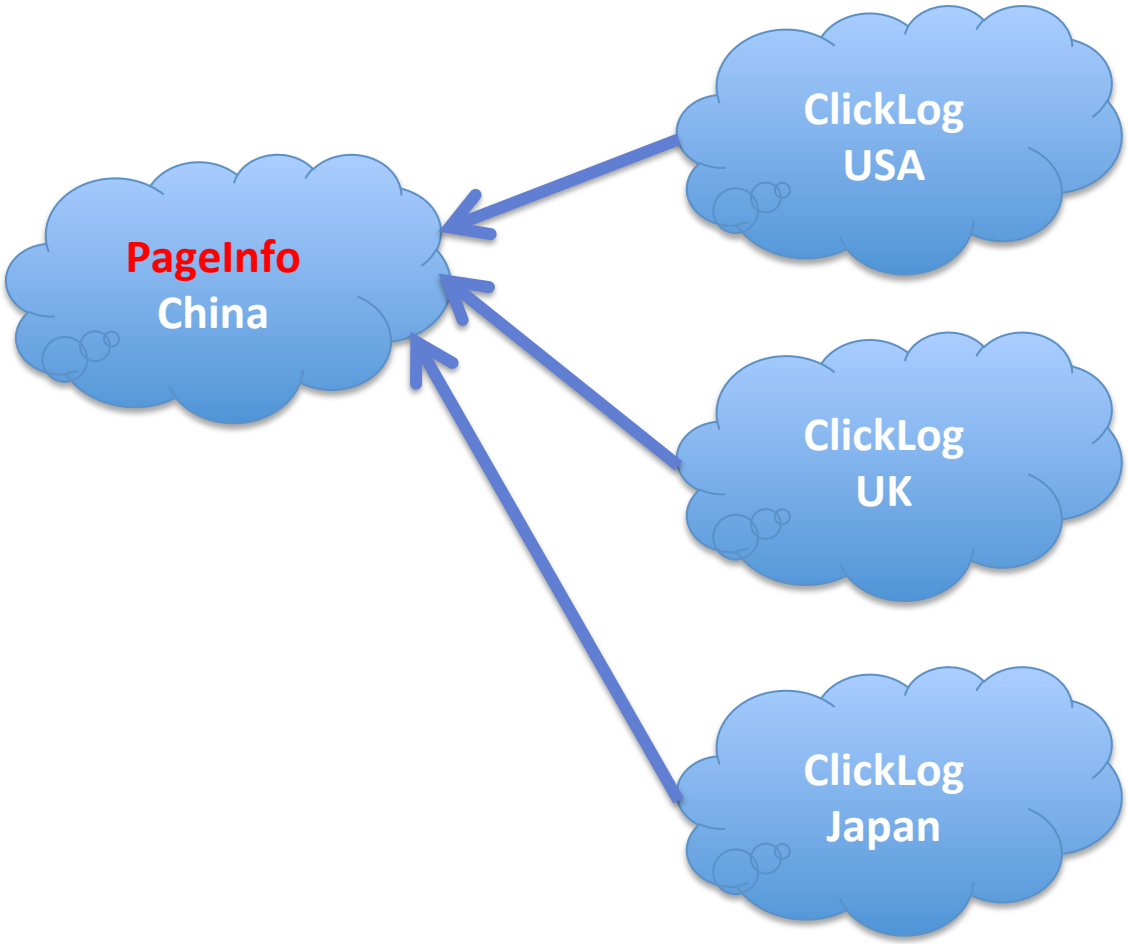
Centralized method



- Copy all ClickLog data to central DC

- Run the query there

Centralized method



- 1B users
- 6 pages visited per user
- 200 bytes per ClickLog row

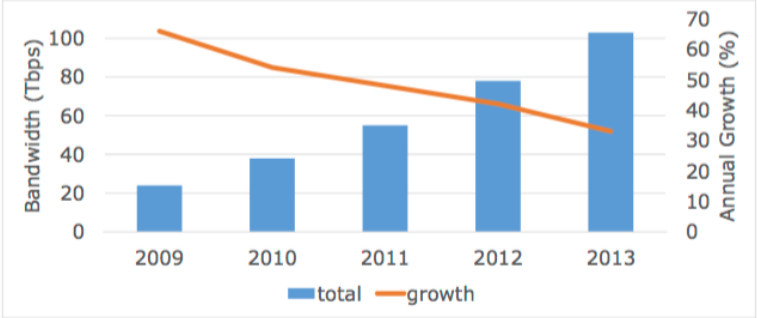
Total data transfer: 1B * 6 * 200 = 1.2 TB

Centralized approach is inadequate

- Consumes scarce, expensive cross-DC bandwidth

Rising costs

Slowing bandwidth growth



Centralized approach is inadequate

- Consumes scarce, expensive cross-DC bandwidth

- Incompatible with sovereignty concerns

  -- Many countries considering restricting moving cities' data

  -- Could render centralized method impossible

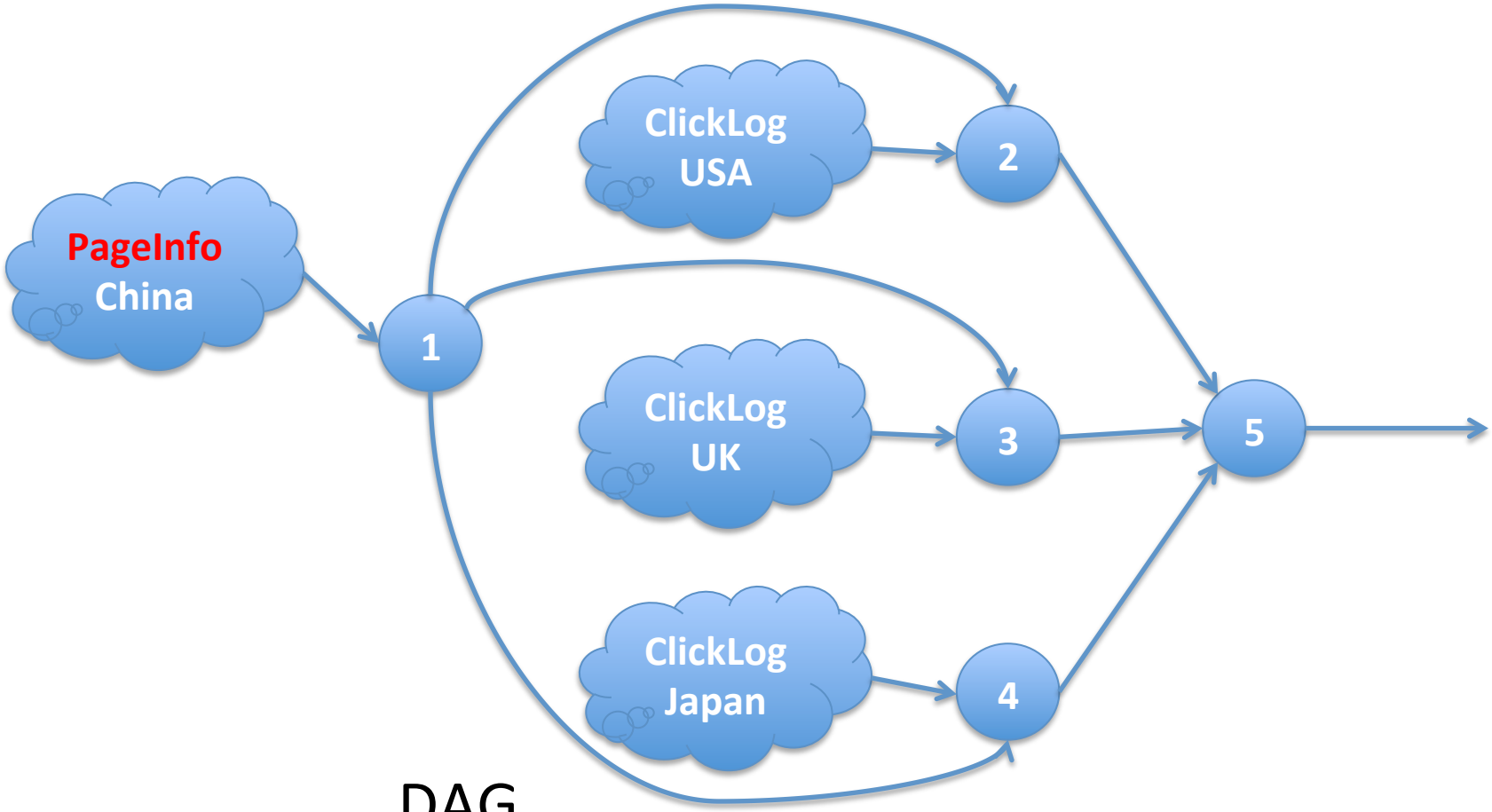
  -- Derived information might still be acceptable

Geo-distributed SQL analytics

PageInfo
(pageURL,
pageSize,
pageRank, …)

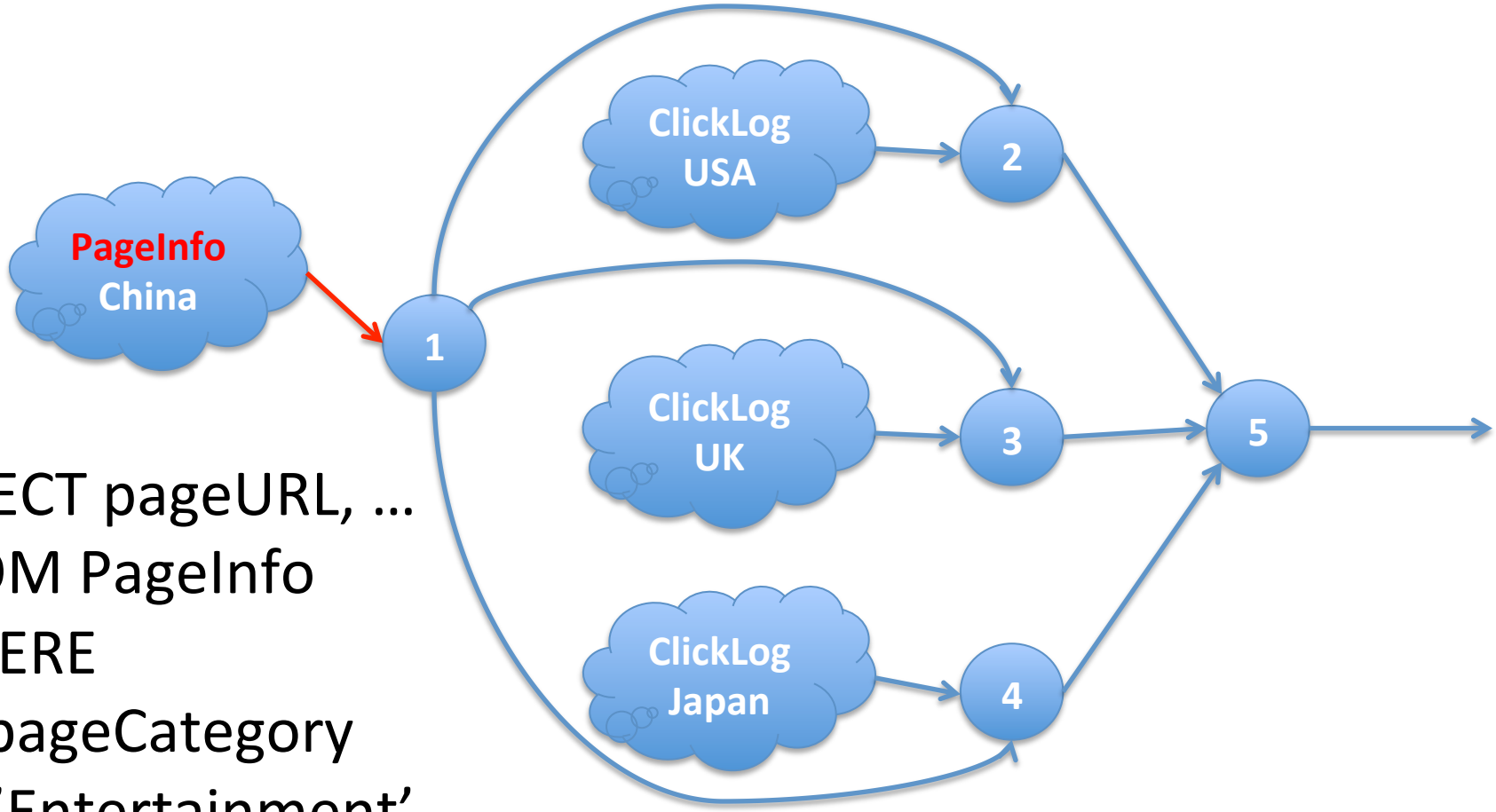
ClickLog
(source IP,
destURL,
visitData,
adRevenue,
…)

Geo-distributed SQL analytics



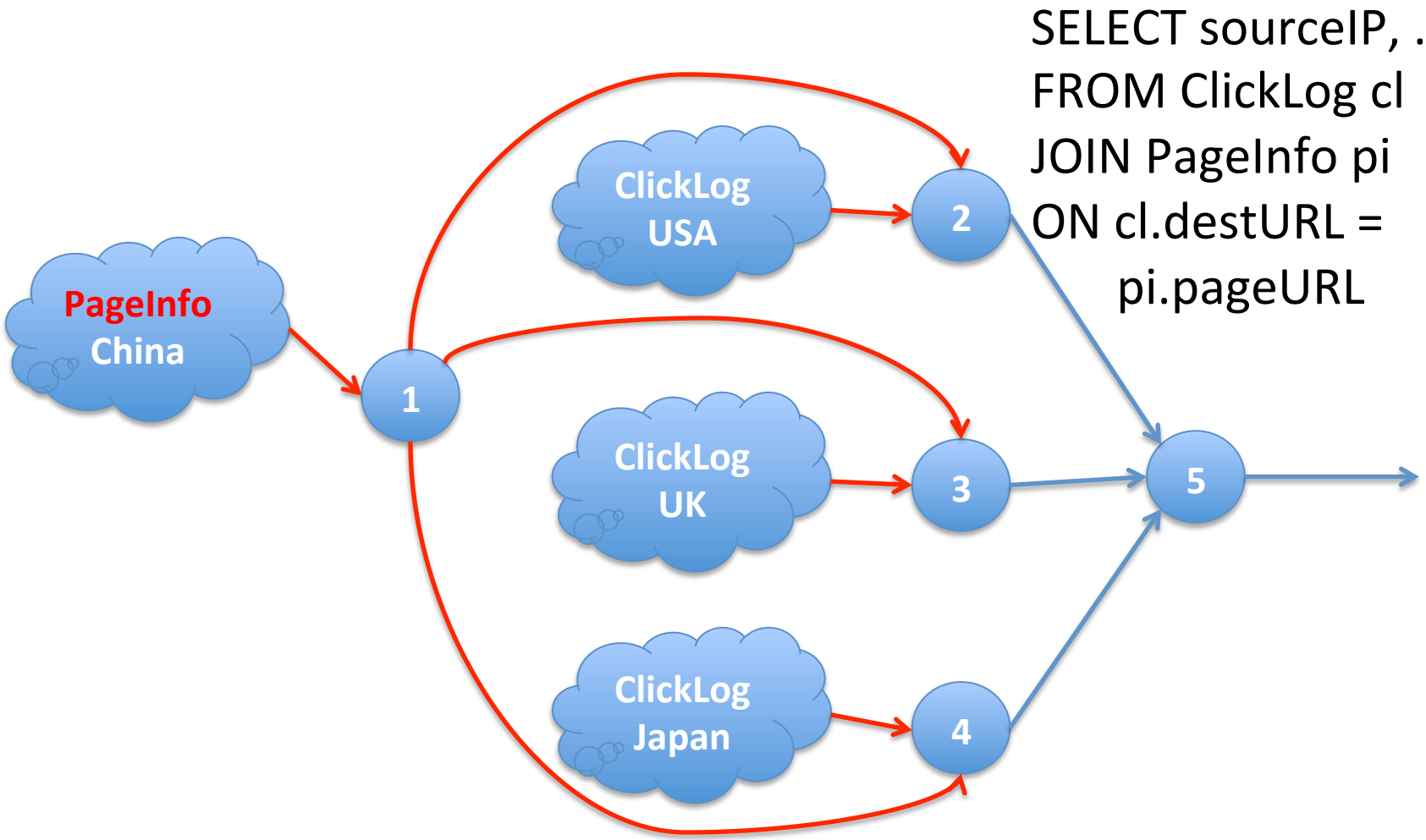
DAG

Geo-distributed SQL analytics

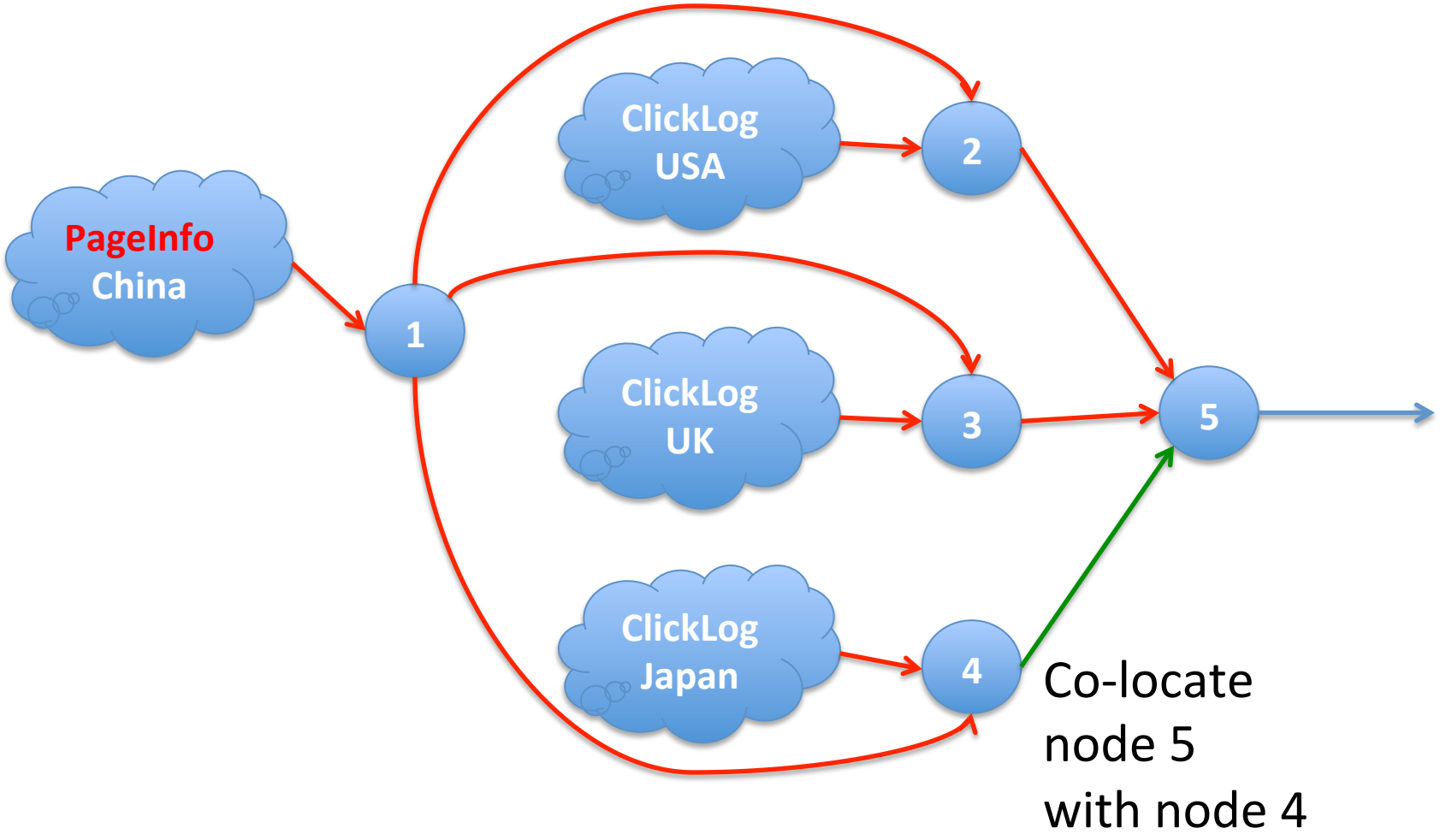



```
SELECT pageURL, …
FROM PageInfo
WHERE
    pageCategory
 =  'Entertainment'
```

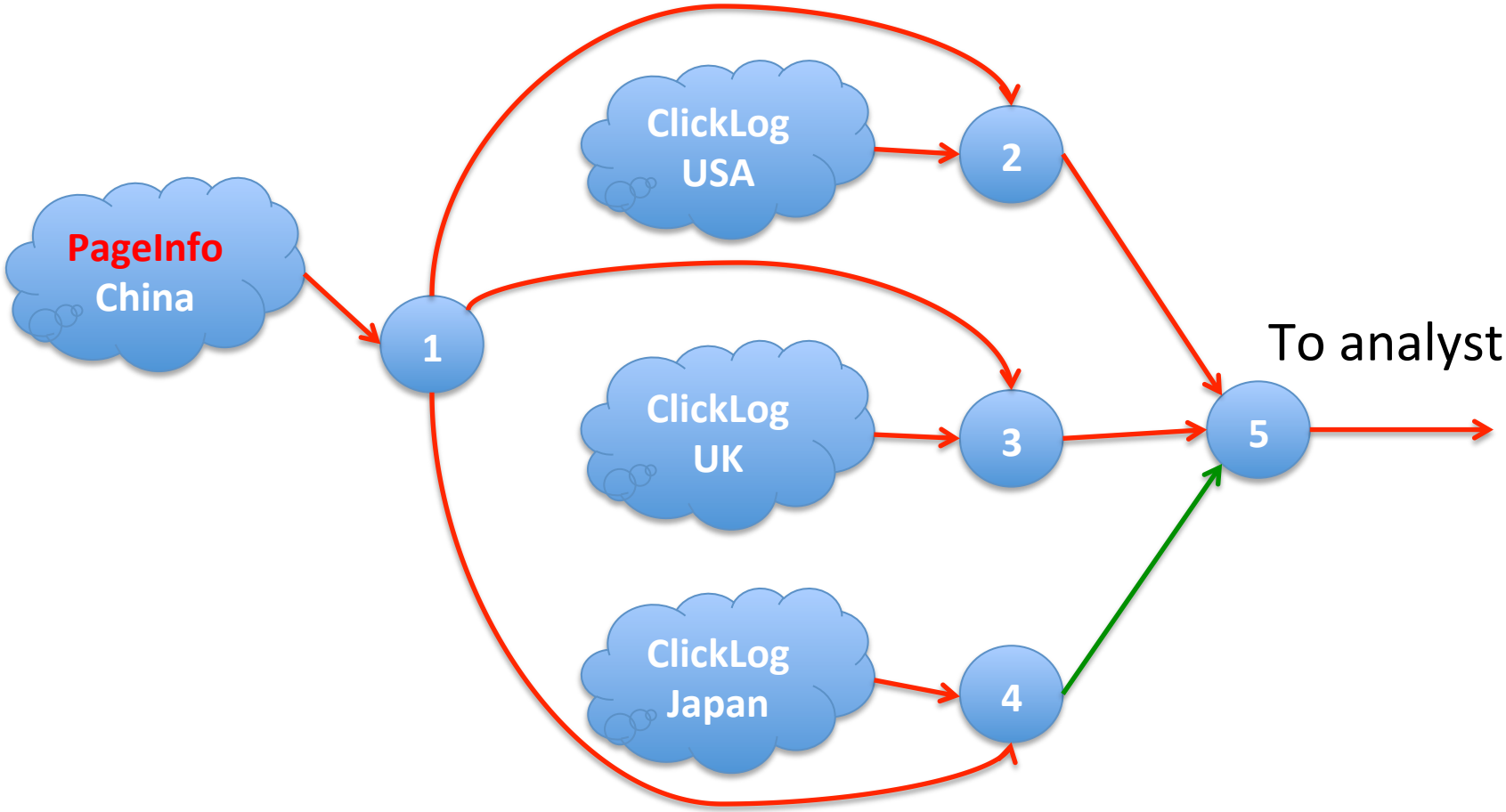
Geo-distributed SQL analytics

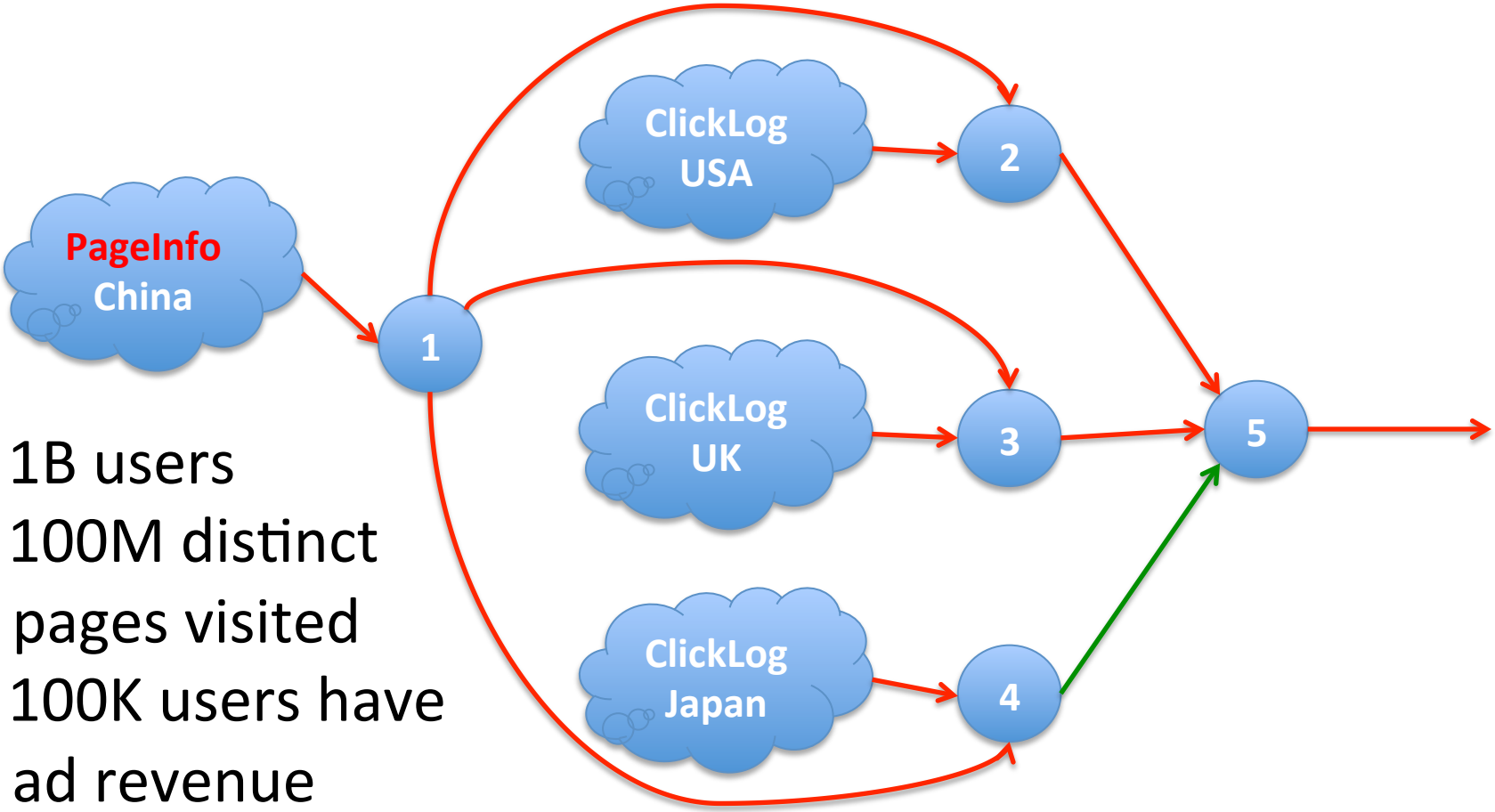


Geo-distributed SQL analytics

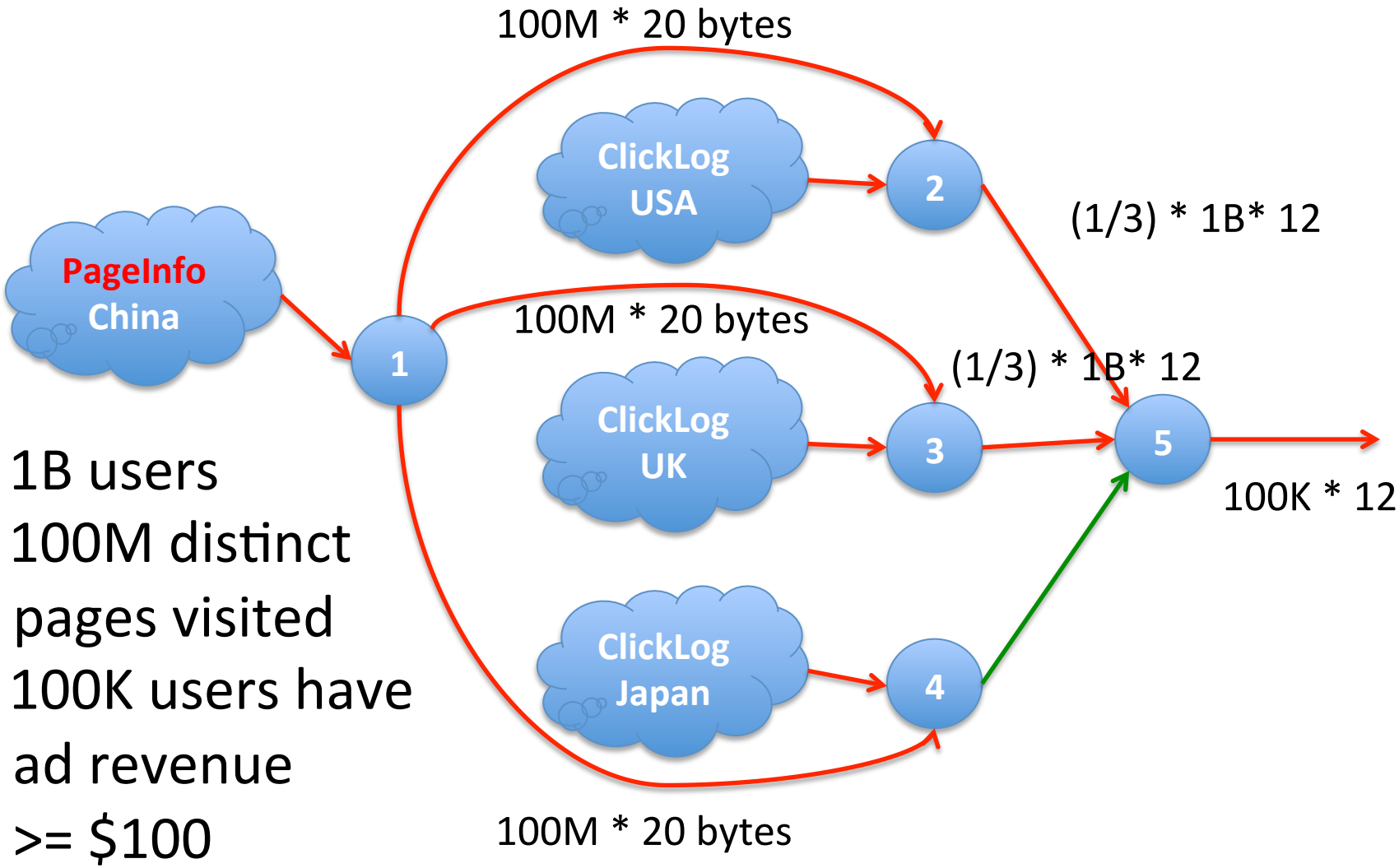


Geo-distributed SQL analytics





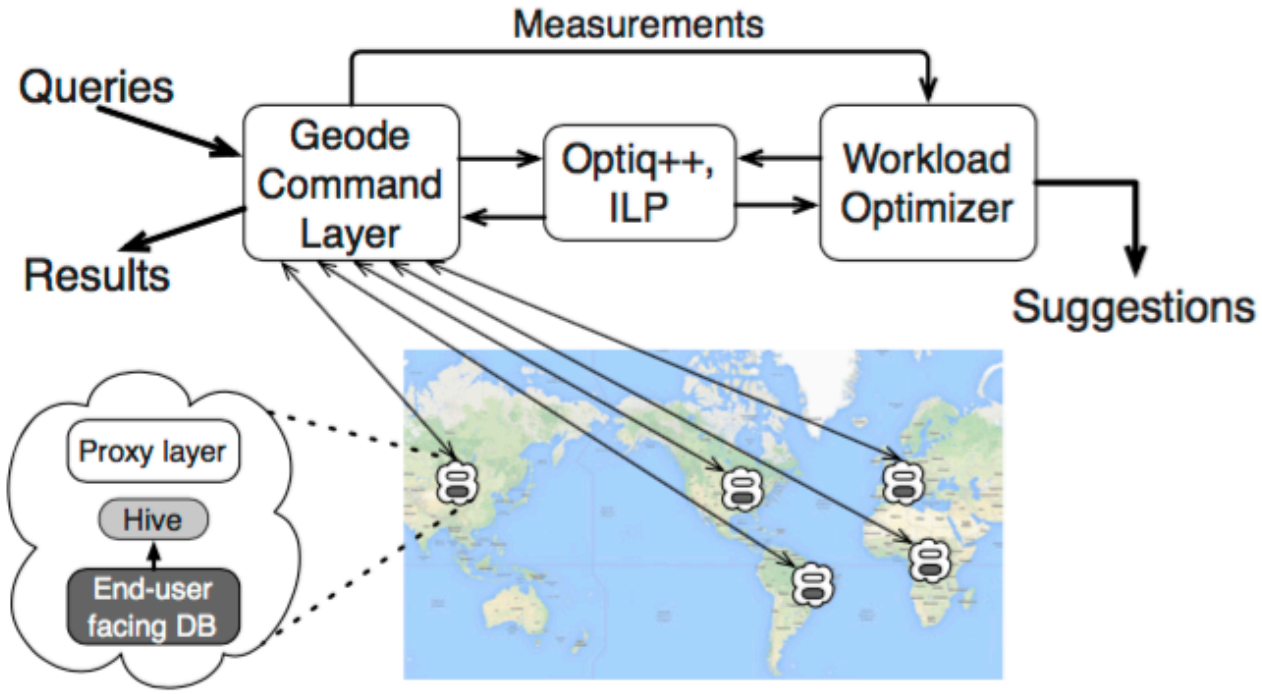
-- 1B users
-- 100M distinct pages visited
-- 100K users have ad revenue >= $100



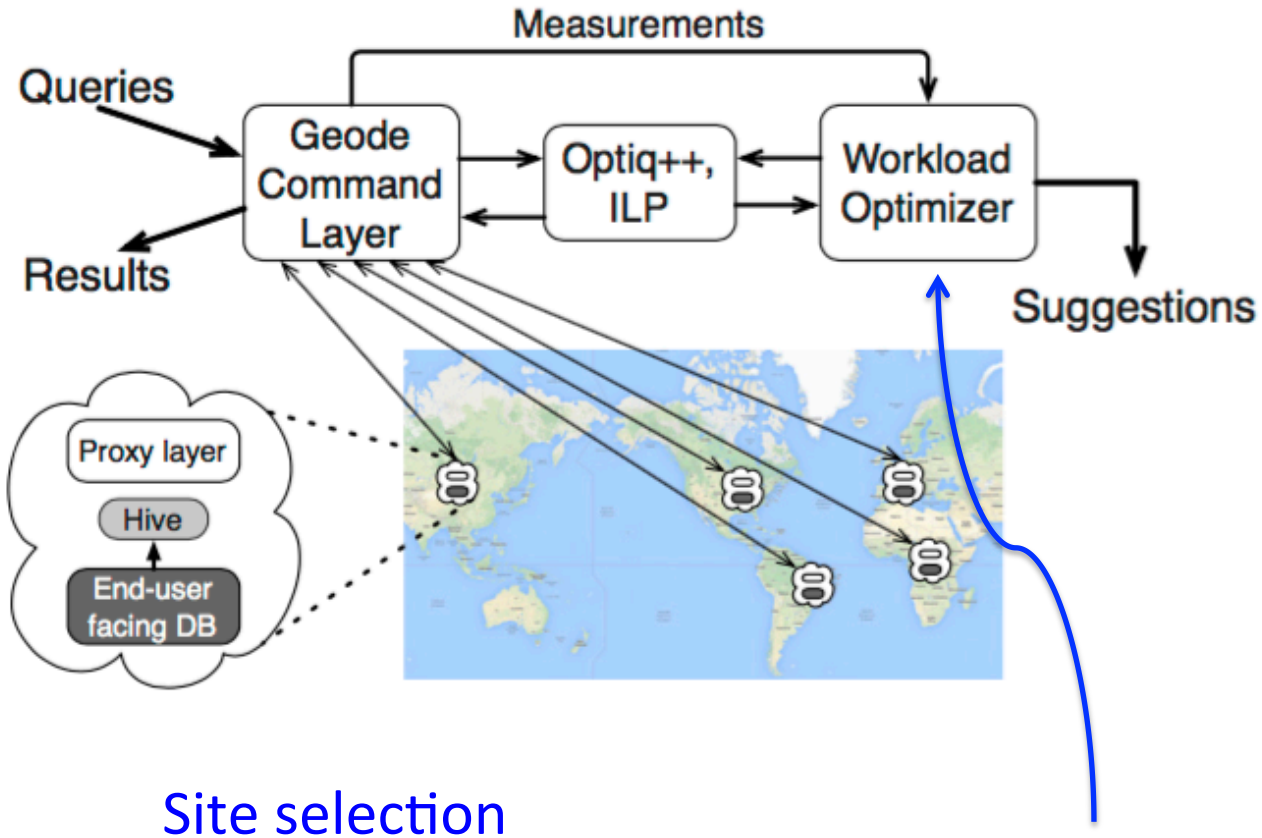
-- 1B users
-- 100M distinct pages visited
-- 100K users have ad revenue >= $100

Total data transfer in inter-DC links:
3 * 100M * 20 + (2/3) * 1B* 12
+ 100K * 12 = 14GB << 1.2TB

Geo-distributed prototype (Geode) architecture



Geode architecture



Site selection

Data (Input data partitions) replication

$$\text{replCost} = \sum_{p=1}^{P} \sum_{d=1}^{D} \text{update\_rate}_p * x_{pd} * \text{link\_cost}_{homeDC(p),d}$$

$$\text{execCost} = \sum_{g \in E} \sum_{d=1}^{D} \sum_{e=1}^{D} y_{gde} * b_g * \text{link\_cost}_{de}$$

copy data partition p from original DC to DC d

transfer data on edge g from DC d to DC e

data volume on edge g

$$\underset{X,Y}{\text{minimize}} \quad \text{replCost} + \text{execCost}$$

subject to

$$\forall (p,d) \in R : x_{pd} = 0$$

$$\forall p : \sum_d x_{pd} \geq f_p$$

$$\forall d \forall e \forall g \mid src(g) \text{ is a partition} : y_{gde} \leq x_{src(g),d}$$

$$\forall d \forall e \forall g \mid src(g) \text{ is a task} : y_{gde} \leq z_{src(g),d}$$

$$\forall n \forall e \forall g \mid dst(g) = n : z_{ne} = \sum_d y_{gde}$$

$$\forall n \forall p \forall d \mid n \text{ reads from partition } p \wedge (p,d) \in R : z_{nd} = 0$$

$$\forall n : \sum_d z_{nd} \geq 1$$

$$\text{replCost} = \sum_{p=1}^{P} \sum_{d=1}^{D} \text{update\_rate}_p * x_{pd} * \text{link\_cost}_{homeDC(p),d}$$

$$\text{execCost} = \sum_{g \in E} \sum_{d=1}^{D} \sum_{e=1}^{D} y_{gde} * b_g * \text{link\_cost}_{de}$$

copy data partition p from original DC to DC d

transfer data on edge g from DC d to DC e

data volume on edge g

$$\underset{X,Y}{\text{minimize}} \quad \text{replCost} + \text{execCost}$$

subject to

$$\forall (p,d) \in R : x_{pd} = 0$$

$$\forall p : \sum_d x_{pd} \geq f_p$$

fault tolerance constraint

$$\forall d \forall e \forall g \mid src(g) \text{ is a partition} : y_{gde} \leq x_{src(g),d}$$

$$\forall d \forall e \forall g \mid src(g) \text{ is a task} : y_{gde} \leq z_{src(g),d}$$

$$\forall n \forall e \forall g \mid dst(g) = n : z_{ne} = \sum_d y_{gde}$$

$$\forall n \forall p \forall d \mid n \text{ reads from partition } p \wedge (p,d) \in R : z_{nd} = 0$$

regulatory constraint

$$\forall n : \sum_d z_{nd} \geq 1$$

Greedy heuristic

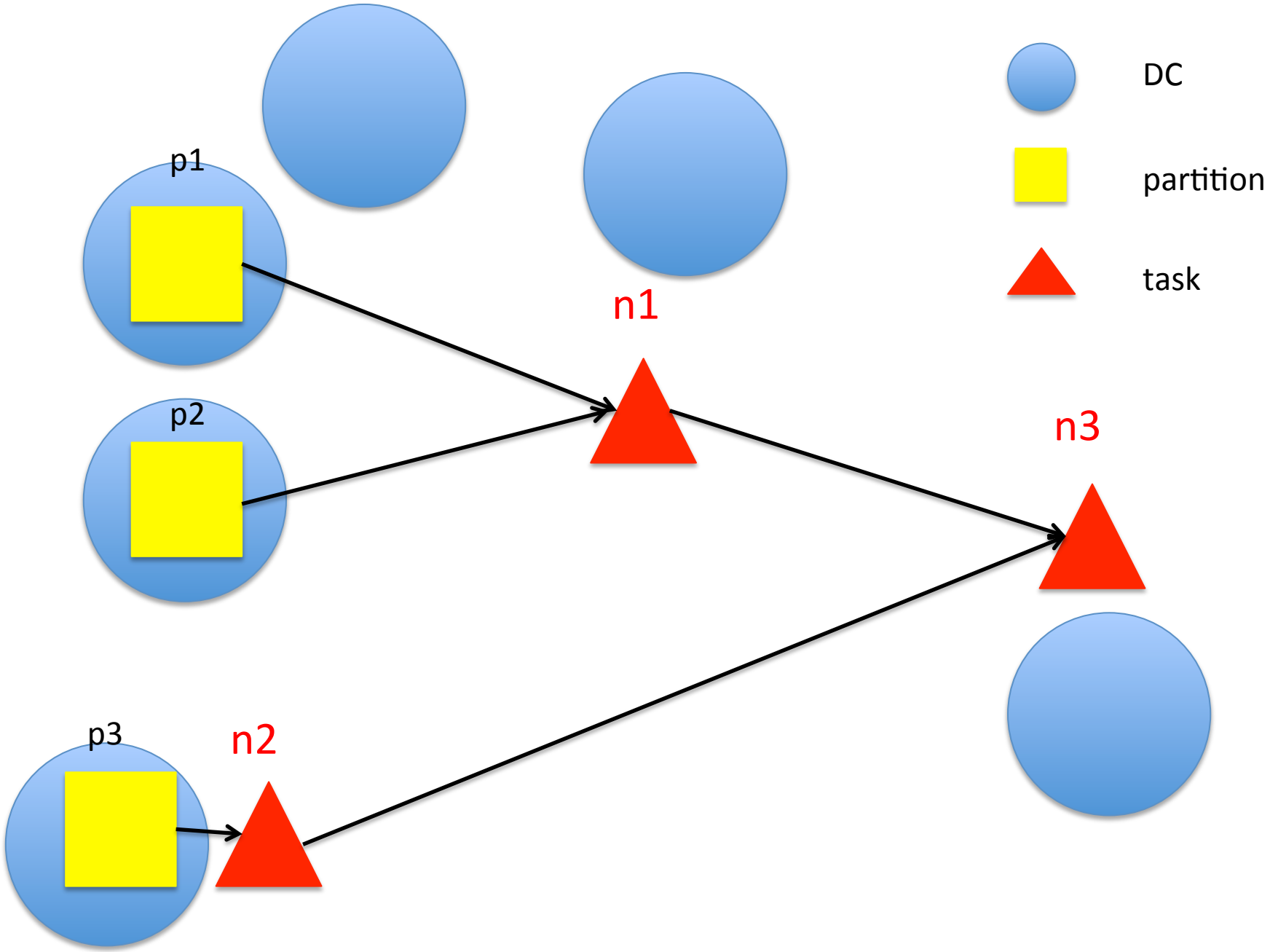
for all DAG $G \in$ workload **do**
    for all task $t \in$ toposort(G) **do**
        for all data center $d \in$ legal_choices(t) **do**
            cost(d) = total cost of copying all of t's inputs to d
        if lowest cost is zero **then**
            assign copies of t to every data center with cost = 0
        else
            assign t to one data center with lowest cost
for all $(p,d) \notin R$ **do**
    check if replicating p to d would further reduce costs
translate decisions so far into values for x, y, z variables in ILP above
solve simplified ILP with pinned values

Greedy heuristic

for all DAG $G \in$ workload **do**
  for all task $t \in$ toposort(G) **do**
    for all data center $d \in$ legal_choices(t) **do**
      cost(d) = total cost of copying all of t's inputs to d
    if lowest cost is zero **then**
      assign copies of t to every data center with cost = 0
    else
      assign t to one data center with lowest cost → each z_{td}
for all $(p,d) \notin R$ **do**
  check if replicating p to d would further reduce costs
translate decisions so far into values for x, y, z variables in ILP above
solve simplified ILP with pinned values

↓

Solve x, y with fixed z and part of y


DC
partition
task
p1
p2
p3
n1
n2
n3

$$\text{replCost} = \sum_{p=1}^{P} \sum_{d=1}^{D} \text{update\_rate}_p * x_{pd} * \text{link\_cost}_{homeDC(p),d}$$

$$\text{execCost} = \sum_{g \in E} \sum_{d=1}^{D} \sum_{e=1}^{D} y_{gde} * b_g * \text{link\_cost}_{de}$$

$$\underset{X,Y}{\text{minimize}} \qquad \text{replCost} + \text{execCost}$$

$x_{homeDC(p)}=1$,
Other $x_{pd}=0$

subject to

$$\forall (p,d) \in R : x_{pd} = 0$$

$$\forall p : \sum_{d} x_{pd} \geq f_p$$

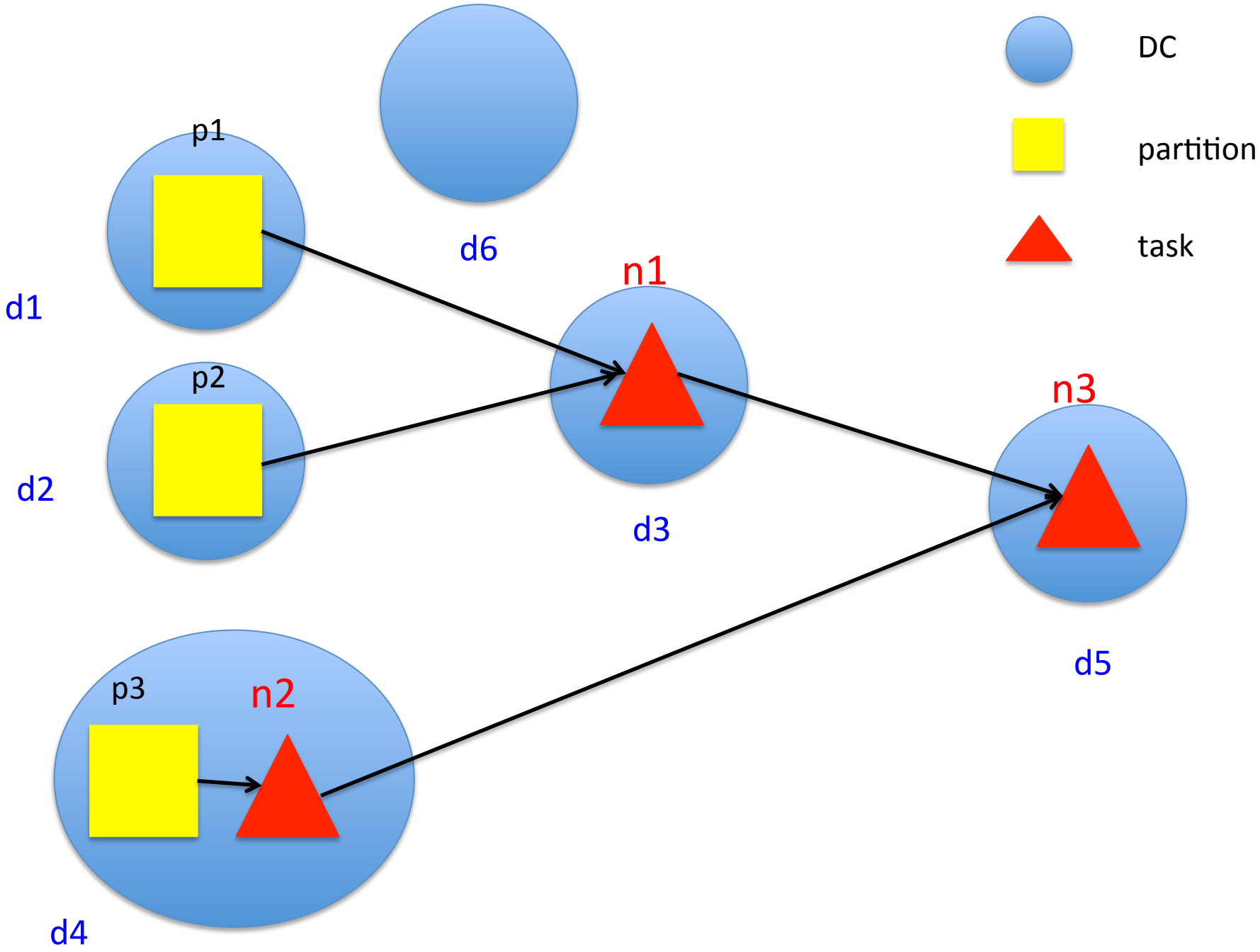
$$\forall d \forall e \forall g \mid src(g) \text{ is a partition} : y_{gde} \leq x_{src(g),d}$$

$$\forall d \forall e \forall g \mid src(g) \text{ is a task} : y_{gde} \leq z_{src(g),d}$$

$$\forall n \forall e \forall g \mid dst(g) = n : z_{ne} = \sum_{d} y_{gde}$$

$$\forall n \forall p \forall d \mid n \text{ reads from partition } p \wedge (p,d) \in R : z_{nd} = 0$$

$$\forall n : \sum_{d} z_{nd} \geq 1$$


DC
partition
task
p1
d1
p2
d2
p3
n2
d4
d6
n1
d3
n3
d5

$$\text{replCost} = \sum_{p=1}^{P} \sum_{d=1}^{D} \text{update\_rate}_p * x_{pd} * \text{link\_cost}_{homeDC(p),d}$$

$$\text{execCost} = \sum_{g \in E} \sum_{d=1}^{D} \sum_{e=1}^{D} y_{gde} * b_g * \text{link\_cost}_{de}$$

$$\underset{X,Y}{\text{minimize}} \qquad \text{replCost} + \text{execCost}$$

subject to

$$\forall (p,d) \in R : x_{pd} = 0$$

$$\forall p : \sum_d x_{pd} \geq f_p$$

$$\forall d \forall e \forall g \mid src(g) \text{ is a partition} : y_{gde} \leq x_{src(g),d}$$

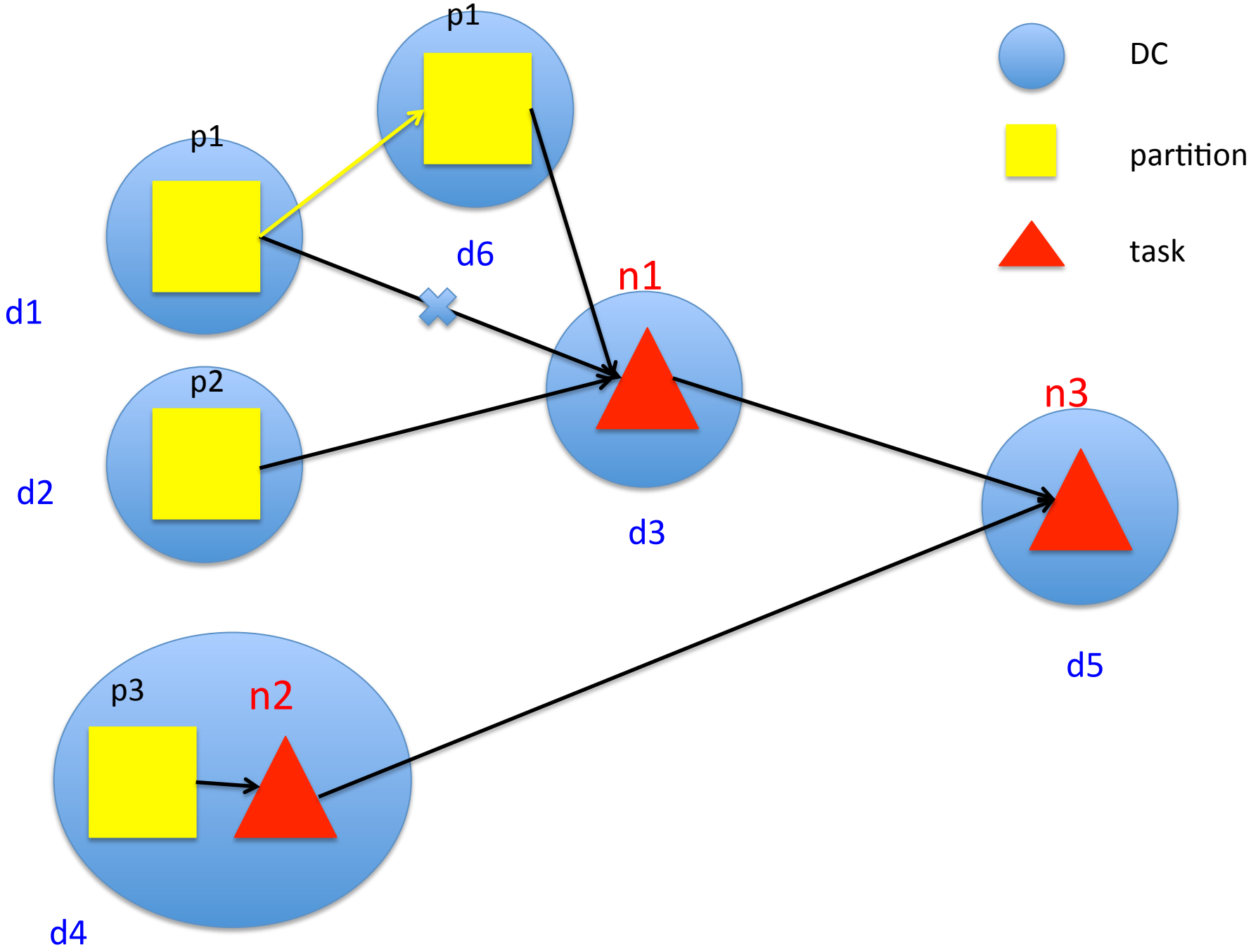
$$\forall d \forall e \forall g \mid src(g) \text{ is a task} : y_{gde} \leq z_{src(g),d}$$

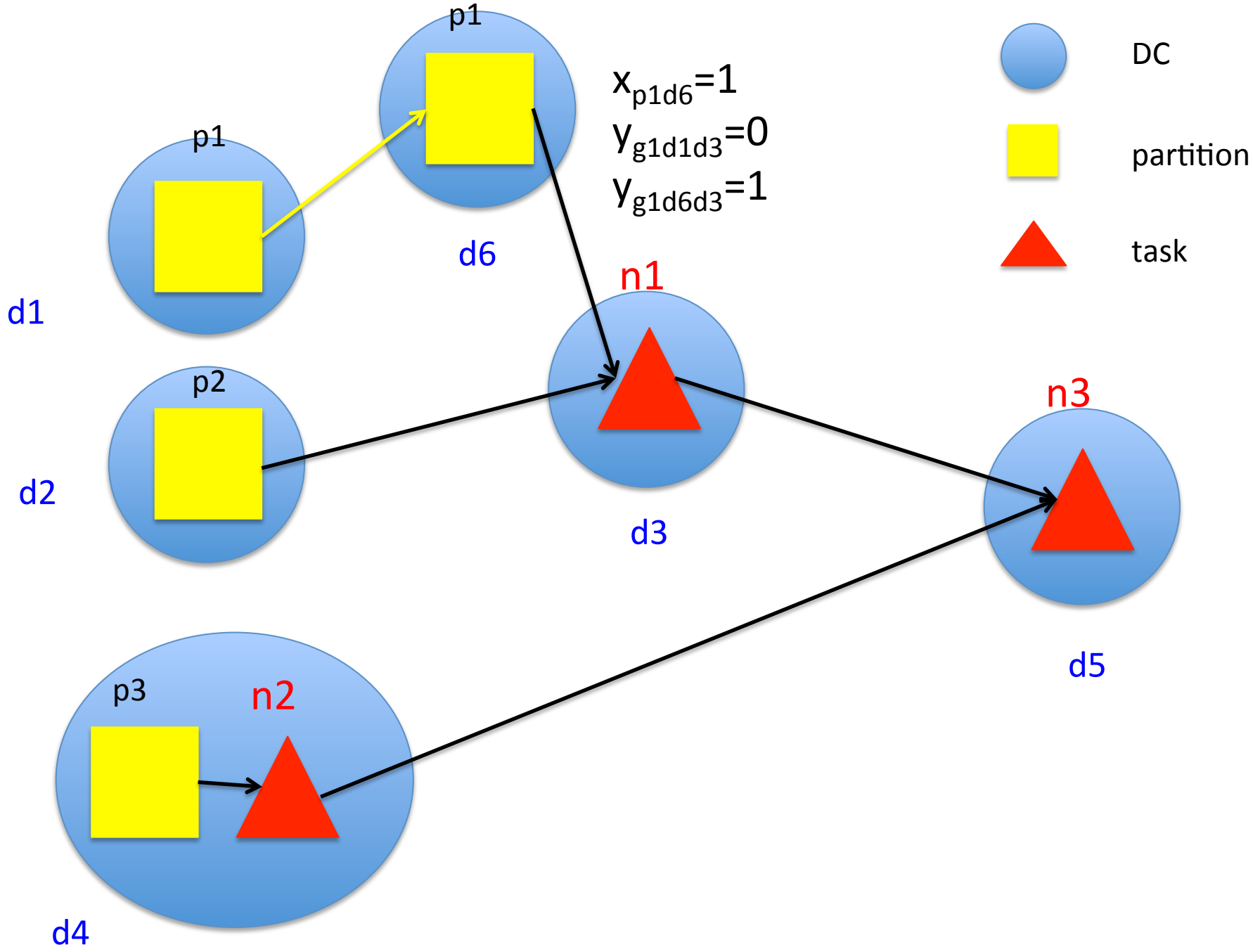
$$\forall n \forall e \forall g \mid dst(g) = n : z_{ne} = \sum_d y_{gde}$$

$$\forall n \forall p \forall d \mid n \text{ reads from partition } p \wedge (p,d) \in R : z_{nd} = 0$$

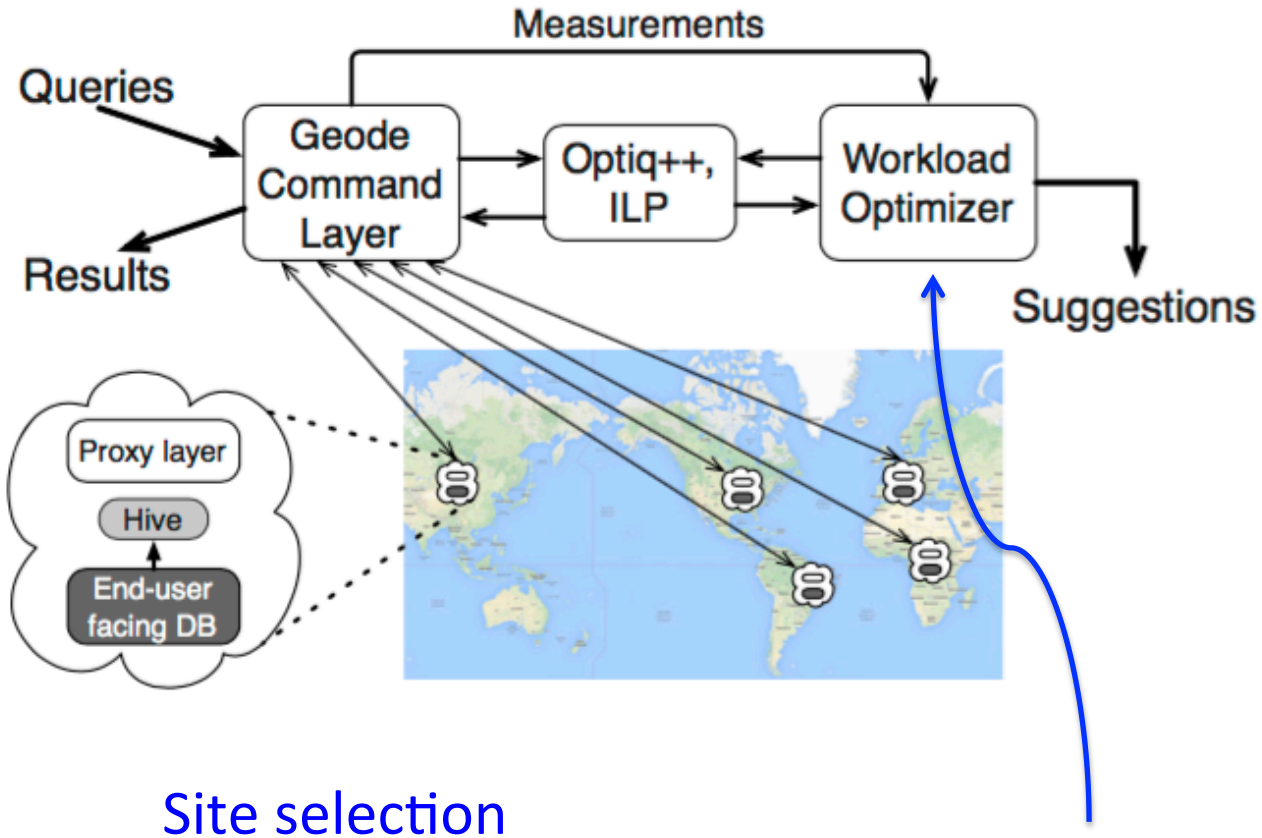
$$\forall n : \sum_d z_{nd} \geq 1$$

fixed when we solve replication decision


p1
p1
d6
d1
n1
p2
d2
d3
n3
d5
p3
n2
d4
DC
partition
task


p1
p1
p1
d1
d6
$x_{p1d6}=1$
$y_{g1d1d3}=0$
$y_{g1d6d3}=1$
DC
partition
task
n1
p2
n3
d2
d3
d5
p3
n2
d4

Geode architecture

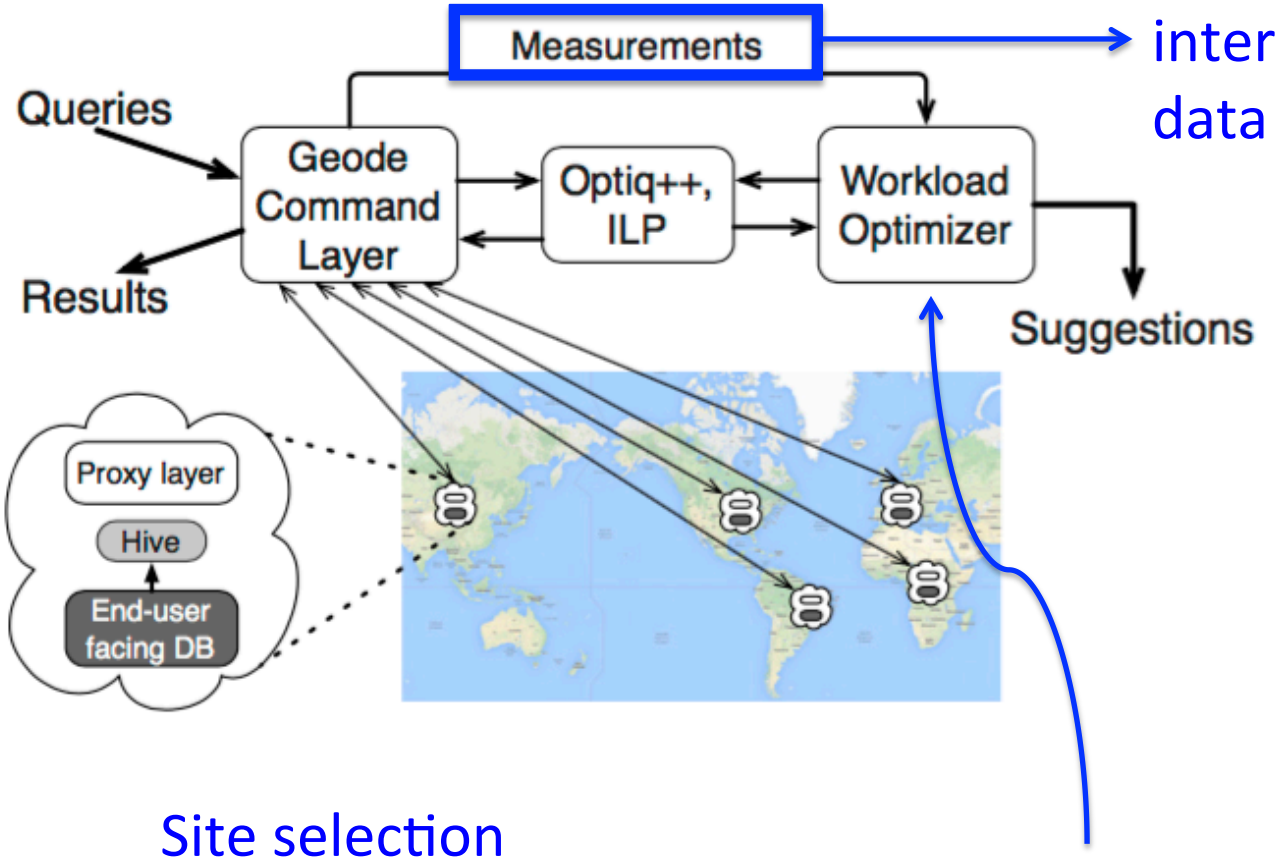


Site selection

Data (Input data partitions) replication

Geode architecture

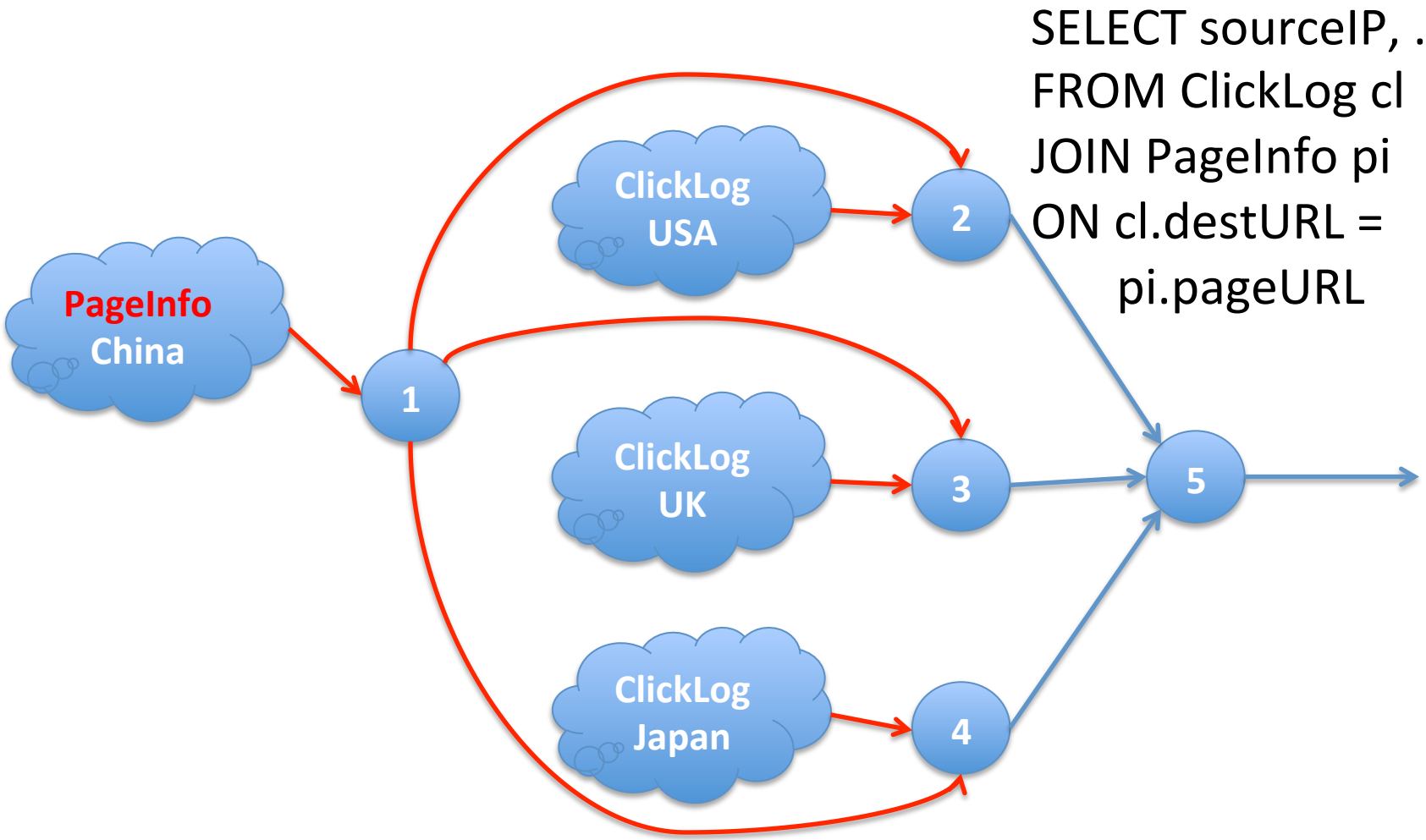
input data volume, intermediate data volume



Site selection

Data (Input data partitions) replication

Geo-distributed SQL analytics

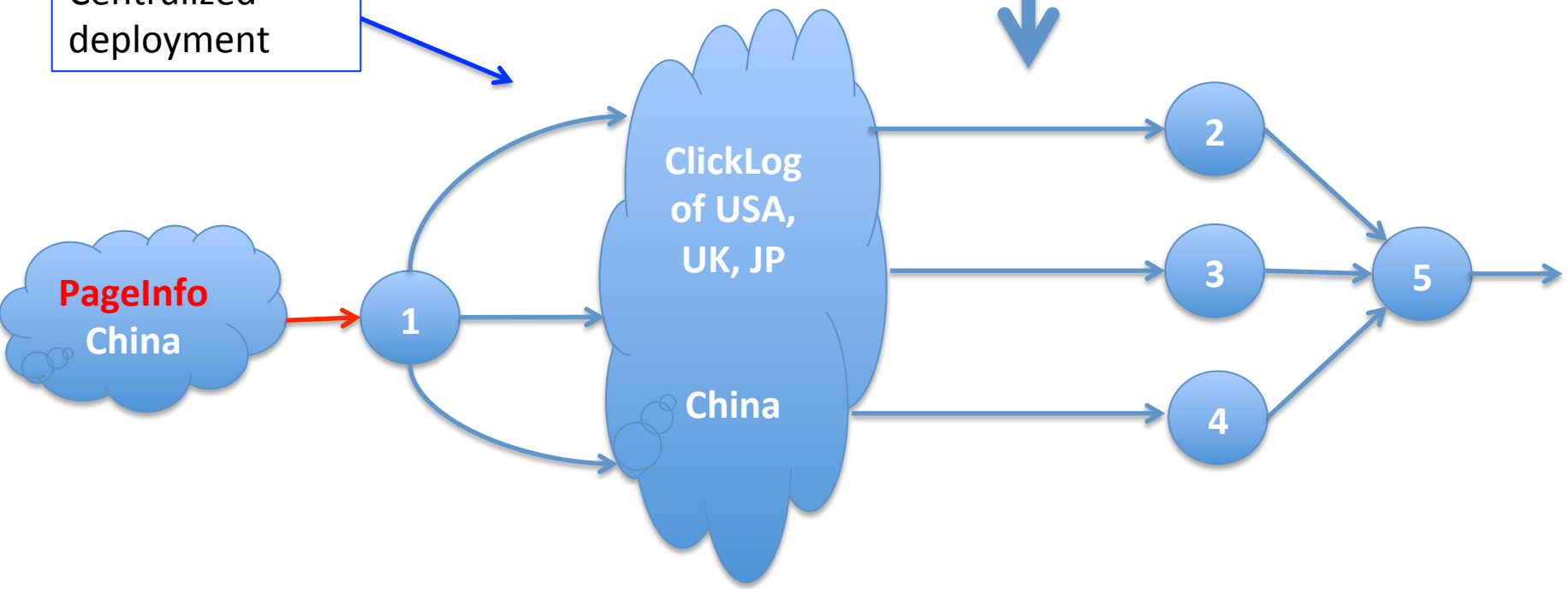


Pseudo-distributed execution

Rewrite queries →

```
SELECT sourceIP, …
WHERE country = X
FROM ClickLog cl
JOIN PageInfo pi ON cl.destURL = pi.pageURL
```

Centralized deployment

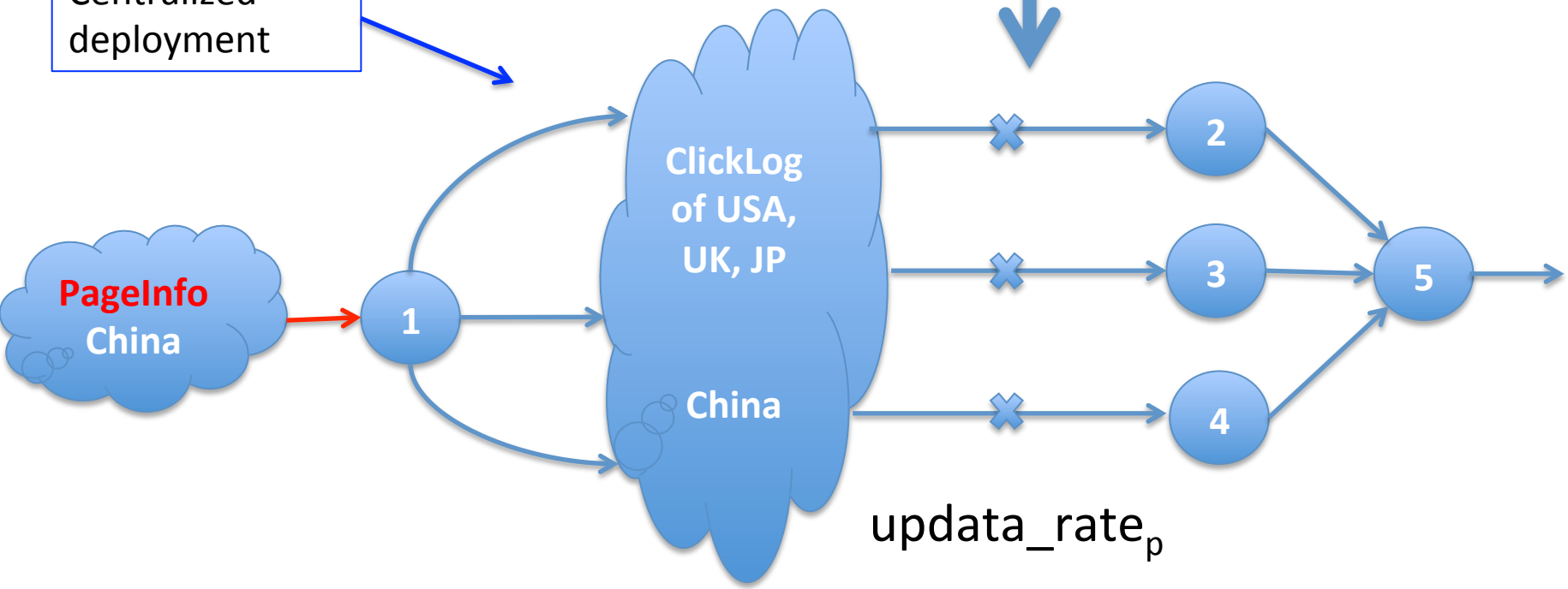


Pseudo-distributed execution

Rewrite queries →

```
SELECT sourceIP, …
WHERE country = X
FROM ClickLog cl
JOIN PageInfo pi ON cl.destURL = pi.pageURL
```

Centralized deployment

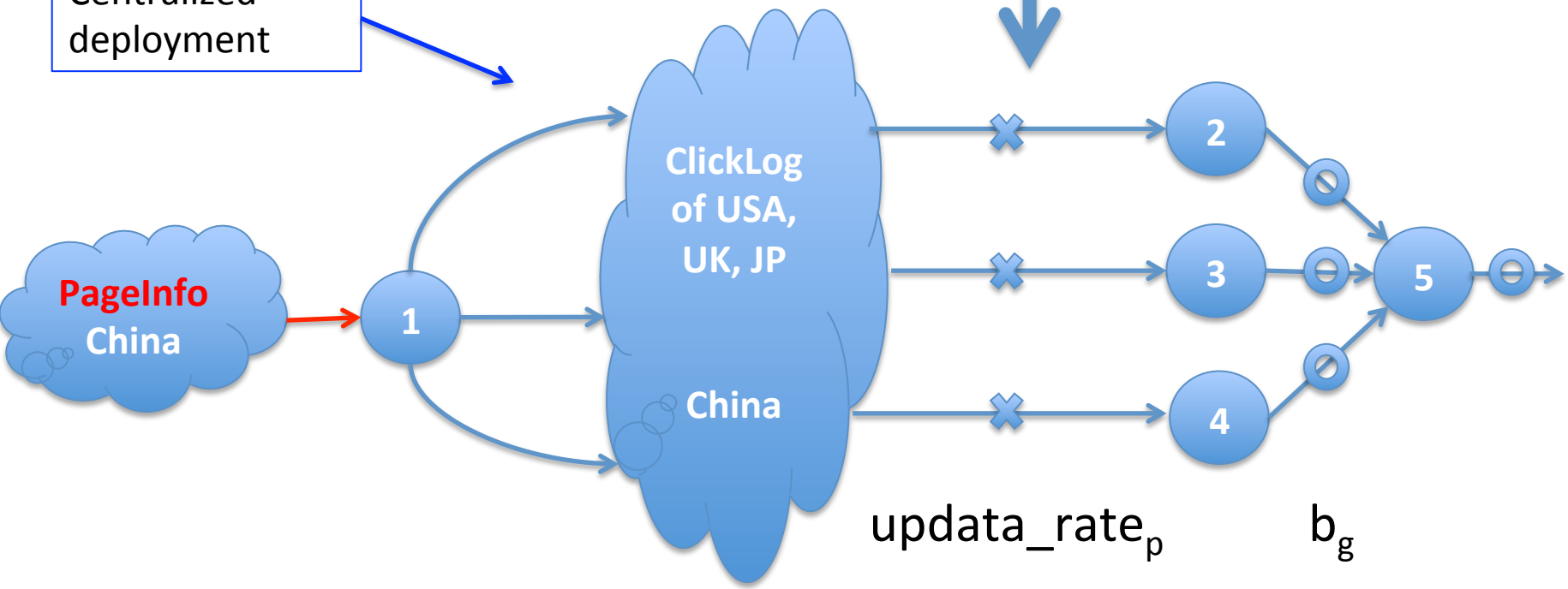


Pseudo-distributed execution

Rewrite queries →

```
SELECT sourceIP, …
WHERE country = X
FROM ClickLog cl
JOIN PageInfo pi ON cl.destURL = pi.pageURL
```

Centralized deployment



input data volume

highly accurate estimations

Intermediate data volume

$$\text{replCost} = \sum_{p=1}^{P} \sum_{d=1}^{D} \text{update\_rate}_p * x_{pd} * \text{link\_cost}_{homeDC(p),d}$$

$$\text{execCost} = \sum_{g \in E} \sum_{d=1}^{D} \sum_{e=1}^{D} y_{gde} * b_g * \text{link\_cost}_{de}$$

$$\underset{X,Y}{\text{minimize}} \quad \text{replCost} + \text{execCost}$$

subject to

$$\forall (p,d) \in R : x_{pd} = 0$$

$$\forall p : \sum_d x_{pd} \geq f_p$$

$$\forall d \forall e \forall g \mid src(g) \text{ is a partition} : y_{gde} \leq x_{src(g),d}$$

$$\forall d \forall e \forall g \mid src(g) \text{ is a task} : y_{gde} \leq z_{src(g),d}$$

$$\forall n \forall e \forall g \mid dst(g) = n : z_{ne} = \sum_d y_{gde}$$

$$\forall n \forall p \forall d \mid n \text{ reads from partition } p \wedge (p,d) \in R : z_{nd} = 0$$

$$\forall n : \sum_d z_{nd} \geq 1$$

Technique contributions

- Execution strategy to minimize bandwidth cost

  -- Classical query planning optimizes join order

- Measurement of data transfer costs across DCs

  -- Classical database techniques are inaccurate

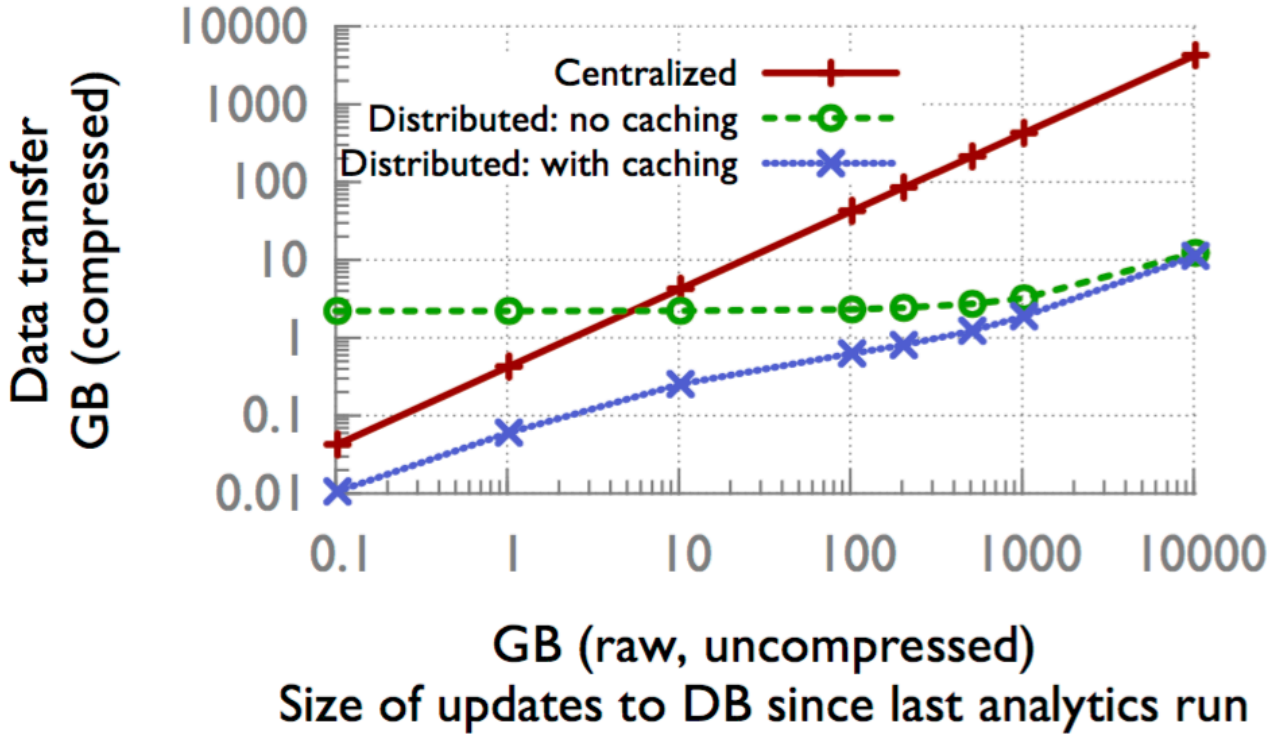
Experimental Evaluation

- How much of the bandwidth saving does the system yield on real workloads?
- What is the tradeoff between solution quality and processing time?
- What is the runtime overhead of collecting the measurements?

Experiment Setup

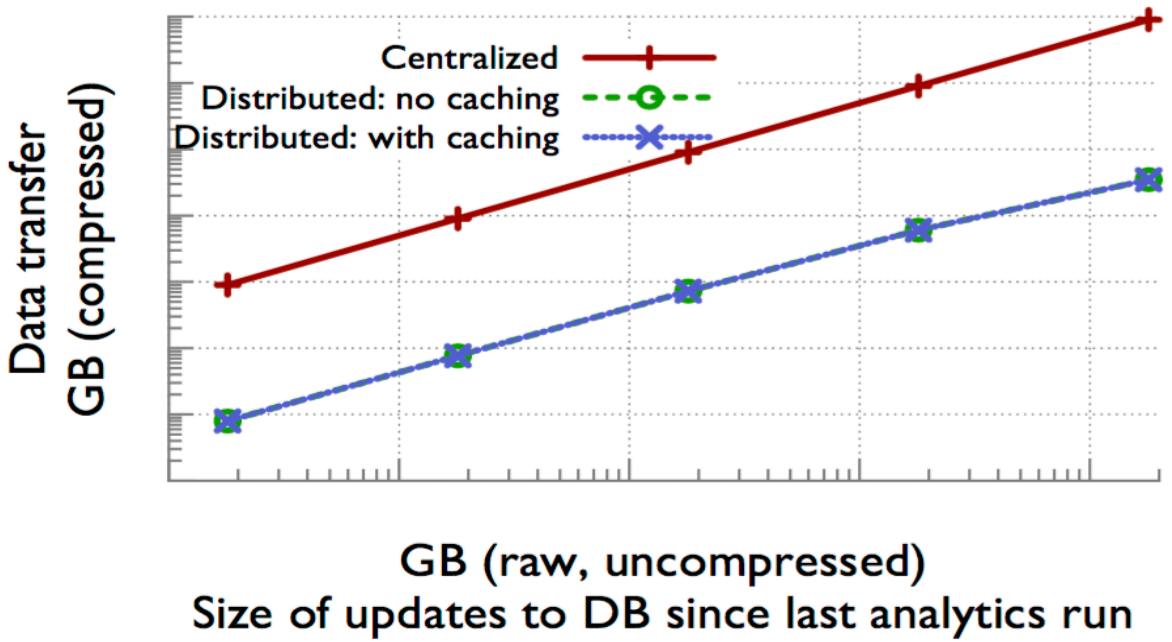
- Centralized VS geo-distributed
  - Three DCs in the US, Europe, and Asia
  - A large centralized cluster
- Input data: 6 workloads
  - E.g., Microsoft production workload
- With Cache VS Without Cache
  - With Cache: cache all the intermediate data
  - Advantage of With Cache: queries are repeated

Bandwidth saving



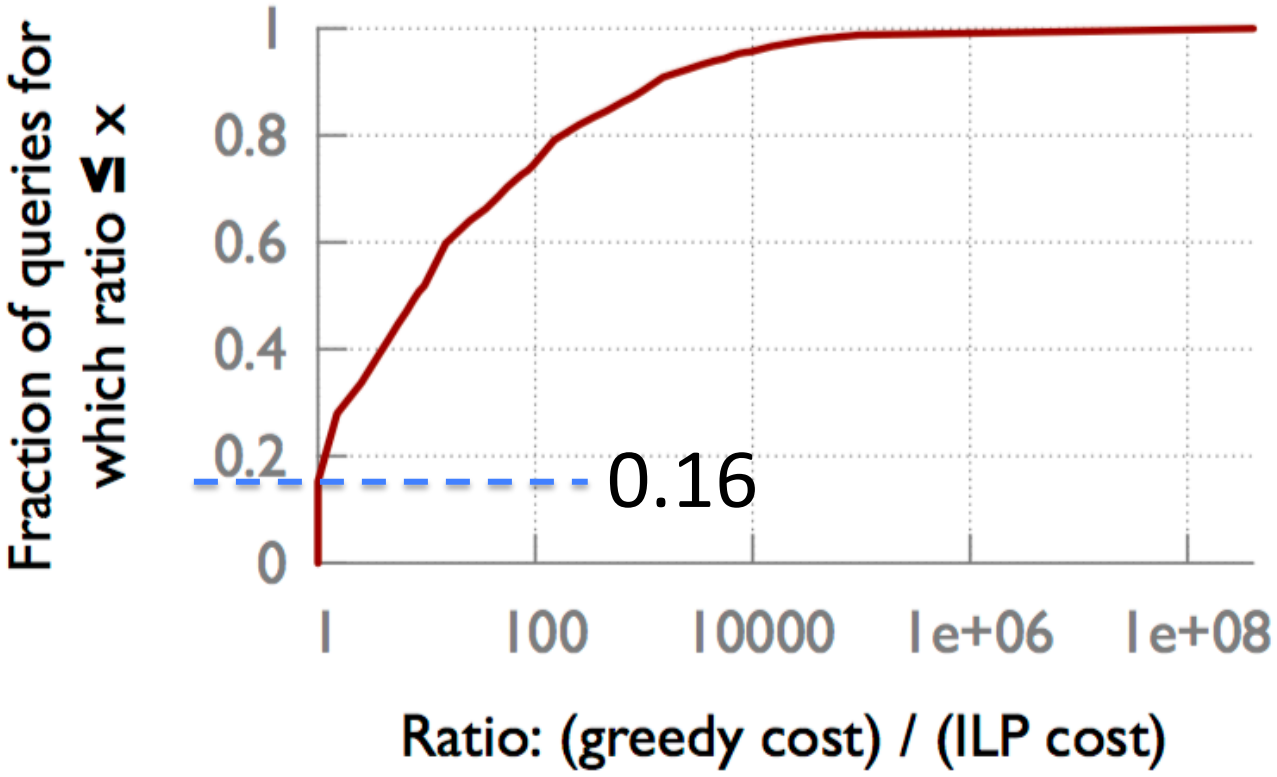
TPC-CH

Bandwidth saving

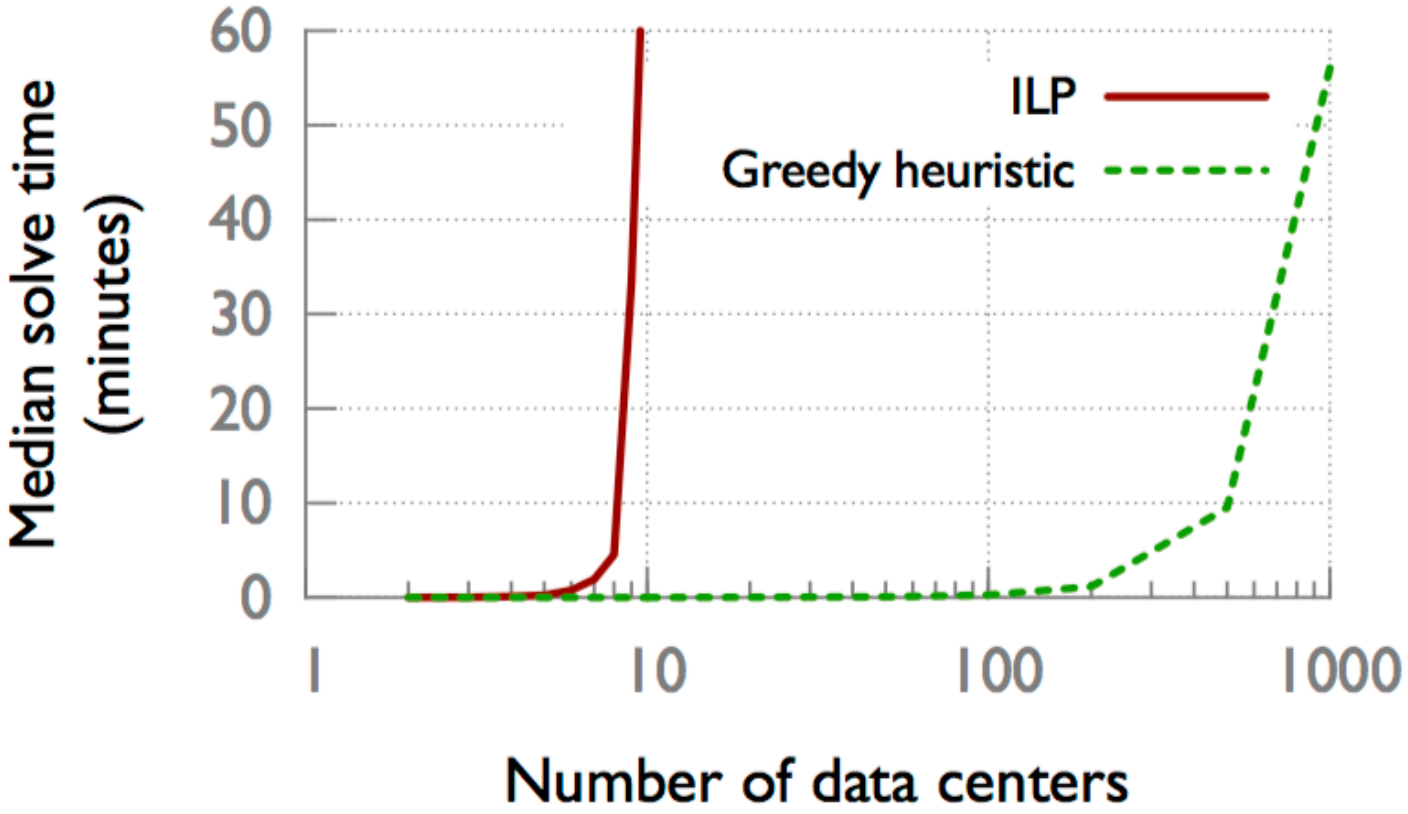


Microsoft production workload

ILP and greedy heuristic comparison



ILP and greedy heuristic comparison



Measurement runtime overhead

- < 20% in runtime

Thank you

Q & A

Potential approximate/online algorithms

- Min-k cut problem (without data replication)
- Uncapacitated facility location problem

$$\text{minimize} \quad \sum_{j=1}^{J} c'_j x_j + \sum_{j=1}^{J}\sum_{i=1}^{I} c_{ij} z_{ij}$$

$$s.t. \quad \sum_{j=1}^{J} z_{ij} \geq 1$$

$$z_{ij} \leq x_j \quad \forall i, j$$

$$x_j, z_{ij} \in \{0, 1\}$$

$$\text{minimize} \quad \sum_{p=1}^{P}\sum_{d=1}^{D} c'_{pd} x_{pd} + \sum_{g \in E}\sum_{d=1}^{D}\sum_{e=1}^{D} c_{gde} y_{gde}$$

$$s.t. \quad \sum_{e=1}^{D} z_{ne} \geq 1 \quad \forall n$$

$$y_{gde} \leq x_{src(g)d} \quad \forall src(g) \in \mathcal{P}, d, e$$

$$y_{gde} \leq z_{src(g),d} \quad \forall src(g) \in \mathcal{N}, e$$

$$z_{ne} = \sum_{d=1}^{D} y_{gde} \quad \forall dst(g) = n, n, e$$