­­BIG DATA ANALYTICS

Week 06 Walkthrough

SQL

1. SHOW DATABASES;
   1. enwiki\_p
   2. dewiki\_p
   3. simplewiki\_p
2. USE simplewiki\_p;
3. SHOW TABLES;
   1. *Wikimedia database layout:* [*https://www.mediawiki.org/wiki/Manual:Database\_layout*](https://www.mediawiki.org/wiki/Manual:Database_layout)
4. DESCRIBE page; #choose any page to have a quick look at the fields it contains.
5. SELECT \* FROM page LIMIT 10;
   1. *page\_namespace:* [*https://www.mediawiki.org/wiki/Manual:Namespace#Built-in\_namespaces*](https://www.mediawiki.org/wiki/Manual:Namespace#Built-in_namespaces)

*### Revision table ###*

1. SELECT \* FROM revision LIMIT 10;

*ORDER the results*

1. SELECT \* FROM revision ORDER by rev\_timestamp LIMIT 10;

*Use DESC, ASC*

1. SELECT rev\_page FROM revision ORDER by rev\_timestamp DESC limit 10;

*JOIN … ON ….*

1. SELECT rev\_page, page\_title, rev\_timestamp FROM revision JOIN page ON rev\_page = page\_id ORDER by rev\_timestamp ASC limit 10;
2. SELECT \* FROM user limit 10;

*double JOIN*

1. SELECT rev\_page, page\_title, rev\_timestamp, rev\_user, user\_name FROM revision JOIN page ON rev\_page = page\_id JOIN user ON rev\_user= user\_id ORDER by rev\_timestamp ASC limit 10;

*COUNT*

1. SELECT COUNT(rev\_id) FROM revision WHERE rev\_user = "66249";

*GROUP*

1. SELECT COUNT(rev\_id) FROM revision GROUP by rev\_user LIMIT 10;
2. SELECT COUNT(rev\_id) AS editcounts FROM revision GROUP by rev\_user ORDER BY editcounts DESC LIMIT 10;
3. SELECT COUNT(rev\_id) AS editcounts, rev\_user, user\_name FROM revision JOIN user ON user\_id = rev\_user GROUP by rev\_user ORDER BY editcounts DESC LIMIT 10;

*WHERE NOT LIKE*

1. SELECT COUNT(rev\_id) AS editcounts, rev\_user, user\_name, user\_registration FROM revision JOIN user ON user\_id = rev\_user WHERE user\_name NOT LIKE "%bot%" AND user\_name NOT LIKE "%Bot%" GROUP by rev\_user ORDER BY editcounts DESC LIMIT 10;

*What is the most edited page by Auntof6?*

1. SELECT COUNT(rev\_id) AS editcounts, page\_title FROM revision JOIN user ON rev\_user = user\_id JOIN page ON page\_id = rev\_page WHERE user\_name="Auntof6" GROUP BY rev\_page ORDER BY editcounts DESC LIMIT 10;

*namespace!*

1. SELECT COUNT(rev\_id) AS editcounts, page\_title FROM revision JOIN user ON rev\_user = user\_id JOIN page ON page\_id = rev\_page WHERE user\_name="Auntof6" and page\_namespace = "0" GROUP BY rev\_page ORDER BY editcounts DESC LIMIT 10;

*Pagelinks*

1. SELECT pl\_from, page\_title FROM pagelinks JOIN page ON page\_id = pl\_from WHERE pl\_title = "Oxford" AND pl\_namespace = "0" LIMIT 10;

**Quiz1: What pages does the page “Oxford” link to?**

1. SELECT pl\_title FROM pagelinks JOIN page ON page\_id = pl\_from WHERE page\_title = "Oxford" AND pl\_namespace = "0" and page\_namespace = "0" LIMIT 10;

**Quiz 2: The most linked to page on Simple Wikipedia?**

1. SELECT COUNT(\*) AS linkscount, pl\_title FROM pagelinks WHERE pl\_namespace = "0" GROUP BY pl\_title ORDER BY linkscount DESC LIMIT 10;

**Quiz 3: What is the title of the page on Simple Wikipedia with the largest number of sister articles in other languages?**

1. SELECT COUNT(\*) AS langcounts, ll\_from, page\_title FROM langlinks JOIN page on ll\_from = page\_id WHERE page\_namespace = 0 GROUP BY ll\_from ORDER BY langcounts DESC LIMIT 10;

*Relation between number of users and number of edits to each page:*

1. *SELECT COUNT(DISTINCT rev\_user) AS usercounts, COUNT(rev\_id) editcounts FROM revision JOIN page ON page\_id = rev\_page group by rev\_page LIMIT 10;*