POSTGRESQL EXECUTOR: EXECUTING YOUR EXECUTION PLAN

RAFIA SABIH

SR. SOFTWARE ENGINEER

CYBERTEC

POSTGRESQL SERVICES & SUPPORT

ABOUT CYBERTEC

Highly specialized,
fast growing
IT company

International Team
(10 countries),
6 locations worldwide

Database , Data &
Science Services

Owner managed
since 2000

DATABASE PRODUCTS & TOOLS

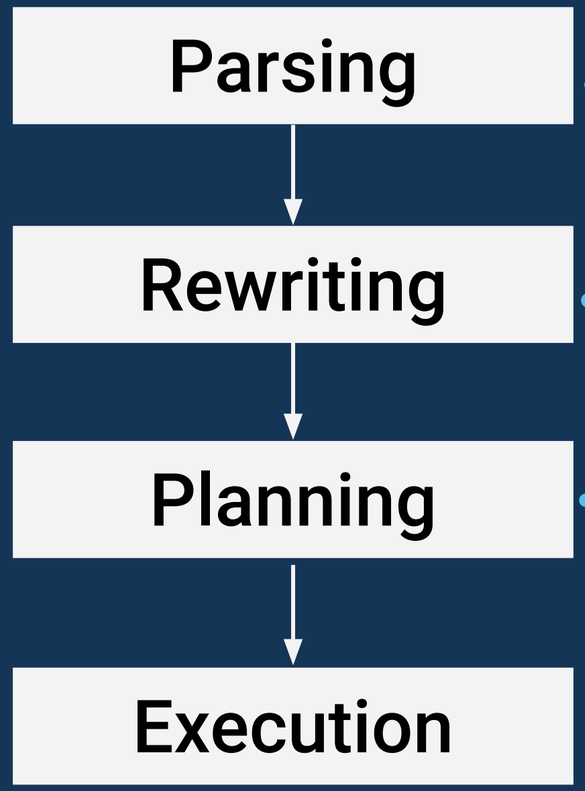
AGENDA

1. CONTROL FLOW OF THE EXECUTOR

2. IMPORTANT DATA STRUCTURES

3. MISCELLANEOUS

POSTGRESQL OVERVIEW



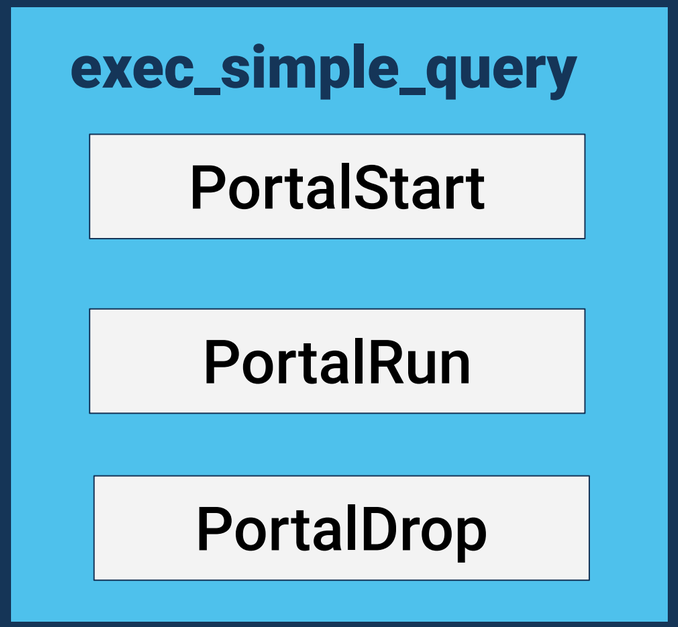
Syntax check, no catalog lookups

Applies rules, rewrite query when using views, etc.

Pick the plan with the lowest cost

JOURNEY OF THE QUERY

- exec_simple_query
  - PortalStart - preparatory phase
  - PortalRun - actual execution
  - PortalDrop - cleanup and close



JOURNEY OF THE QUERY

- Portal (defined in portal.h)
  - `active snapshot`
  - `queryDesc`
  - `sub transaction information`
  - `parameters to pass to query`
  - `Portal strategy - select, update, etc.`

Portals are an abstraction for the execution state of the query

JOURNEY OF THE QUERY

PREPARATION PHASE

- PortalStart
  - ExecutorStart
    - standard_ExecutorStart
    - if there is any function for this hook, then it runs now
- standard_ExecutorStart
  - takes queryDesc as an input
  - tupDesc is now filled to describe the returning tuples

JOURNEY OF THE QUERY

PREPARATION PHASE

- QueryDesc (defined in execdesc.h)
  - `snapshot to be used for the query`
  - `tupDesc`
  - `Estate`
  - `Planstate`
  - `total time spent in query execution`

QueryDesc encapsulates everything required by the executor

JOURNEY OF THE QUERY

PREPARATION PHASE

- Estate (defined in execnodes.h)
  - `Nodetag`
  - `ScanDirection`
  - `List of range tables in query`
  - `Index relations`
  - `relations`
  - `parameters info - internal, external`
  - `memory context`
  - `dsa_area - required for parallel query`

Working state for an executor invocation

JOURNEY OF THE QUERY

PREPARATION PHASE

- standard_ExecutorStart
  - Create Executor EState
  - switches into per query memory context
  - InitPlan
    - ExecInitNode
      - Calls the init function for the respective Plan Node
      - ExecInitAgg, ExecInitSeqScan

JOURNEY OF THE QUERY

EXECUTION PHASE

- PortalRun
  - ExecutorRun
    - standard_ExecutorRun
    - if any hooks are installed, that code runs now
    - ExecutePlan
      - ExecProcNode
        - Keeps on executing the planstate node, till the number of tuples required is reached

JOURNEY OF A QUERY

EXECUTION: SELECT QUERY

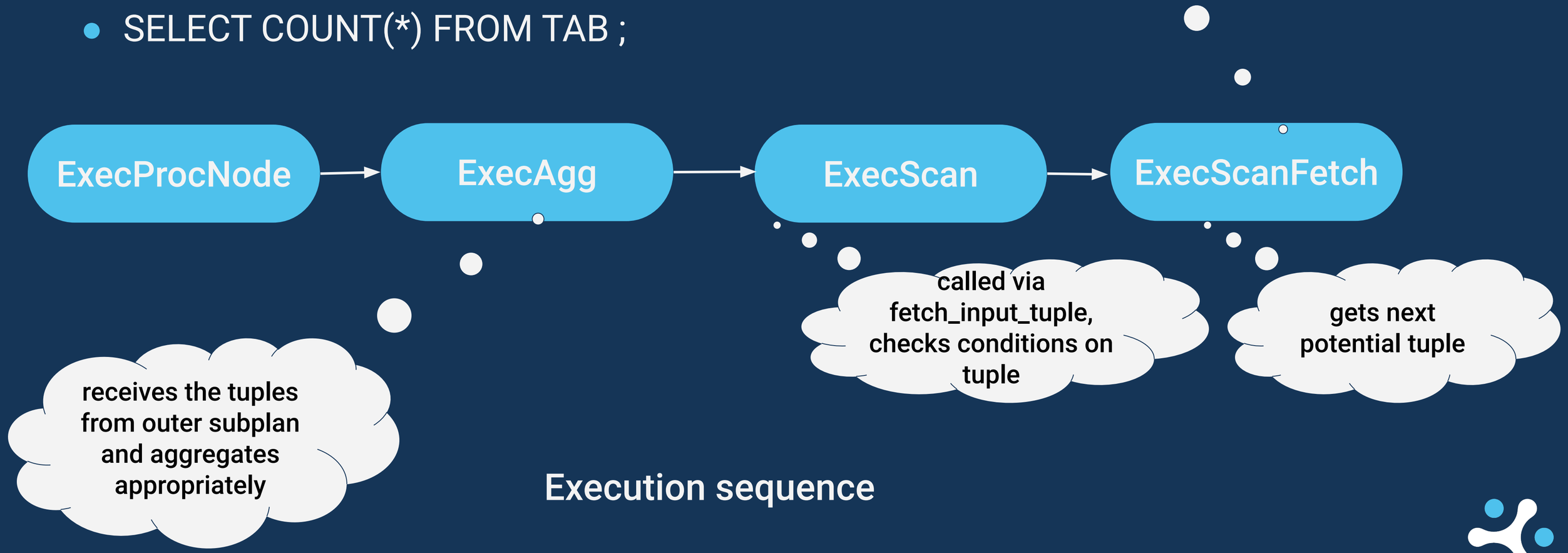
- SELECT COUNT(*) FROM TAB ;

```
1| Aggregate  (cost=163004.04..163004.05 rows=1 width=8) (actual time=505.401..505.401
   rows=1 loops=1)
2| ->  Seq Scan on tab  (cost=0.00..139255.63 rows=9499363 width=0) (actual
   time=0.143..291.517 rows=9502608 loops=1)
3| Planning Time: 0.313 ms
4| Execution Time: 505.474 ms
5| (4 rows)
```

JOURNEY OF A QUERY

EXECUTION: SELECT QUERY

- SELECT COUNT(*) FROM TAB ;



JOURNEY OF A QUERY

EXECUTION: JOINS

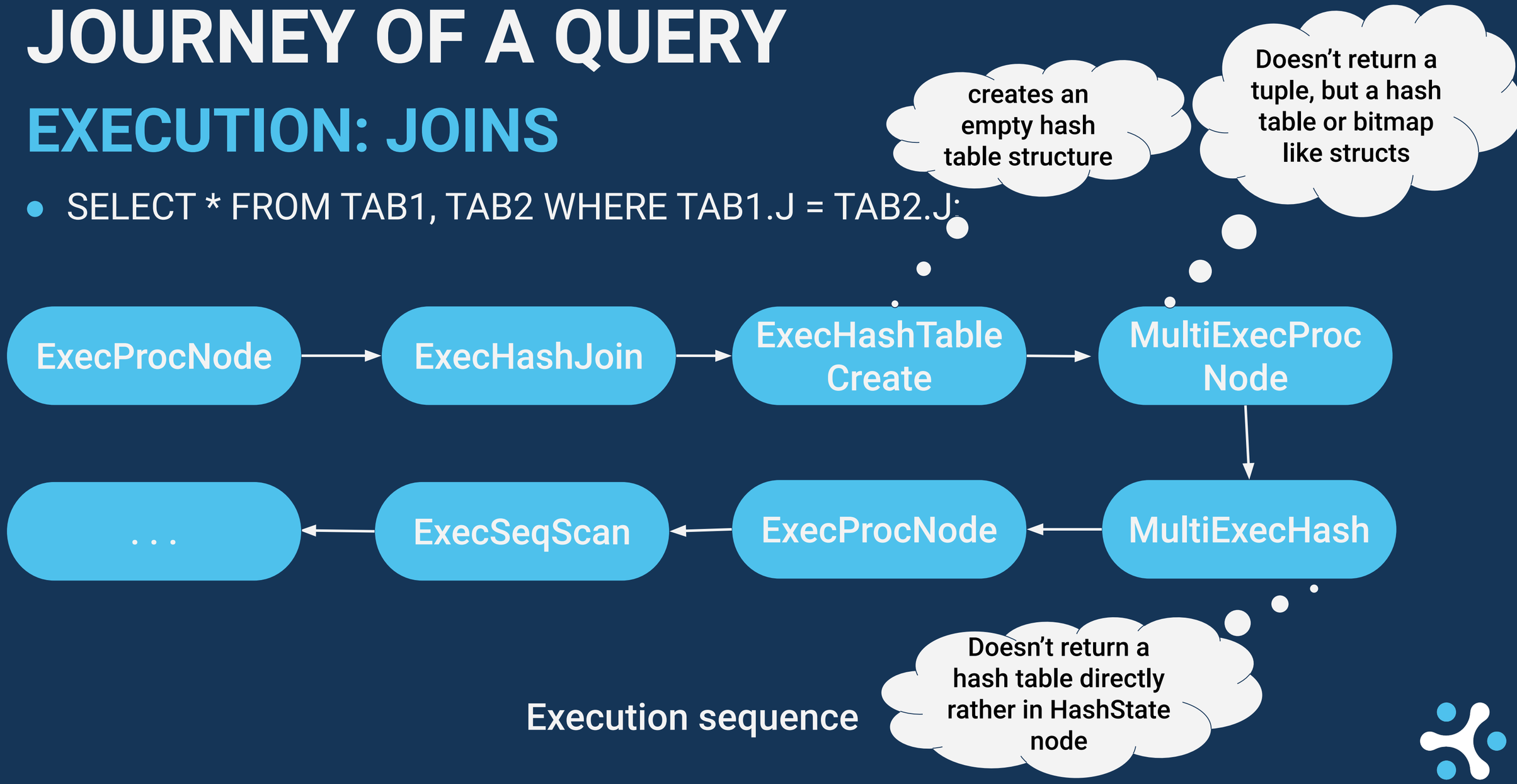
- SELECT * FROM TAB1, TAB2 WHERE TAB1.J = TAB2.J;

```
1| Hash Join  (cost=78.25..174985.49 rows=2900 width=16) (actual time=1.224..676.400
   rows=5800 loops=1)
2|   Hash Cond: (tab.j = tab2.j)
3|   ->  Seq Scan on tab  (cost=0.00..139255.63 rows=9499363 width=8) (actual
     time=0.148.. rows=...loops=1)
4|   ->  Hash  (cost=42.00..42.00 rows=2900 width=8) (actual time=0.907..0.907 rows=2900
     loops=1)
5|         Buckets: 4096  Batches: 1  Memory Usage: 138kB
6|         ->  Seq Scan on tab2  (cost=0.00..42.00 rows=2900 width=8) (actual
          time=0.112..0.380 rows=... loops=1)
7| Planning Time: 0.402 ms
8| Execution Time: 676.821 ms
9| (8 rows)
```

JOURNEY OF A QUERY

EXECUTION: JOINS

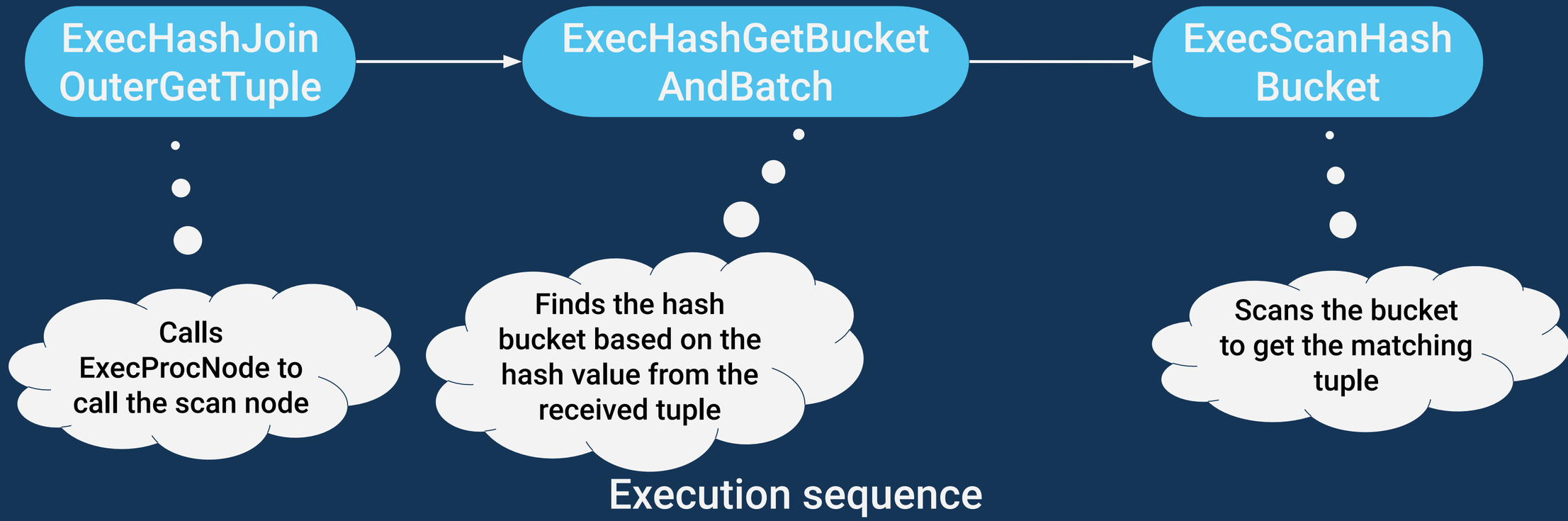
- SELECT * FROM TAB1, TAB2 WHERE TAB1.J = TAB2.J;



JOURNEY OF A QUERY

EXECUTION: JOINS

- SELECT * FROM TAB1, TAB2 WHERE TAB1.J = TAB2.J;



JOURNEY OF A QUERY

EXECUTION: INSERT

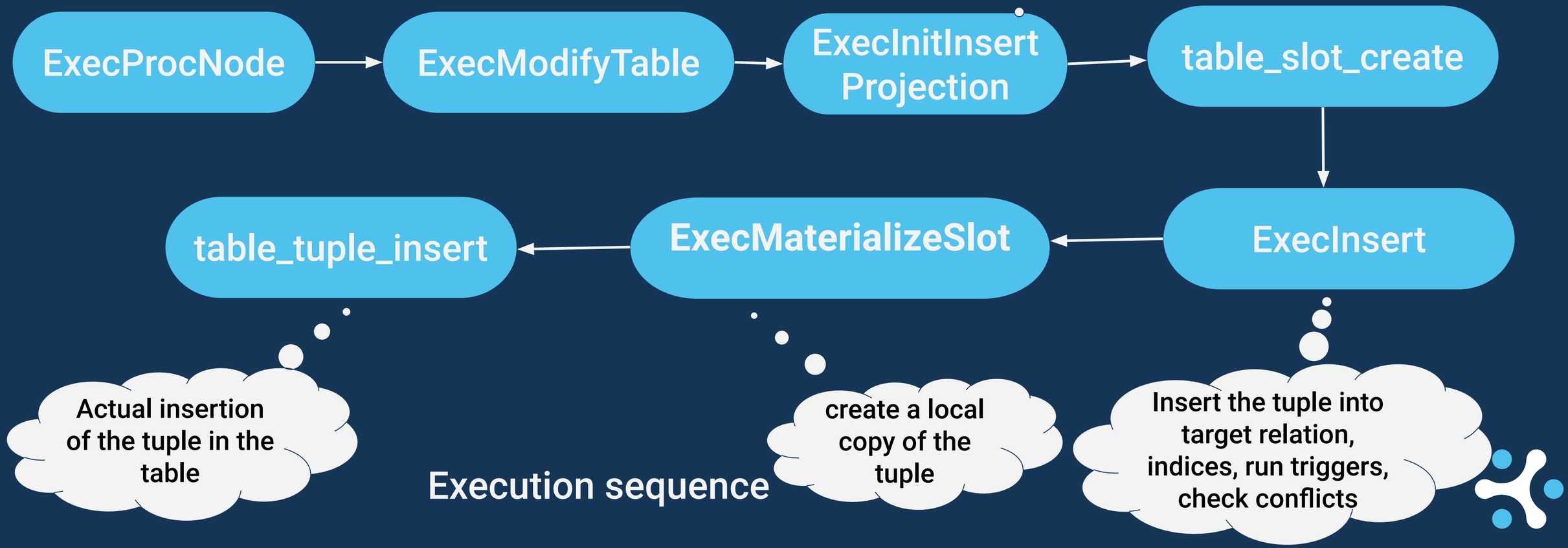
- INSERT INTO TAB VALUES (1,2);

```
1| Insert on tab  (cost=0.00..0.01 rows=0 width=0) (actual time=0.707..0.708 rows=0
   loops=1)
2|  ->  Result  (cost=0.00..0.01 rows=1 width=8) (actual time=0.137..0.138 rows=1
   loops=1)
3| Planning Time: 0.146 ms
4| Execution Time: 0.758 ms
5| (4 rows)
```

JOURNEY OF A QUERY

EXECUTION: INSERT

- INSERT INTO TAB VALUES (1,2);



JOURNEY OF THE QUERY

EXECUTION: PARALLEL QUERY

- Parallel Dynamic shared memory
- ParallelContext
  - `Maximum number of workers to launch`
  - `nworkers_launched`
  - `*error_context_stack`
  - `dsm_segment *seg;`
- TupleQueueReader
  - A DestReceiver of type DestTupleQueue, which is a TQueueDestReceiver writes tuples from the executor to a shm_mq
  - A TupleQueueReader reads tuples from a shm_mq and returns the tuples

JOURNEY OF THE QUERY

EXECUTION: PARALLEL QUERY

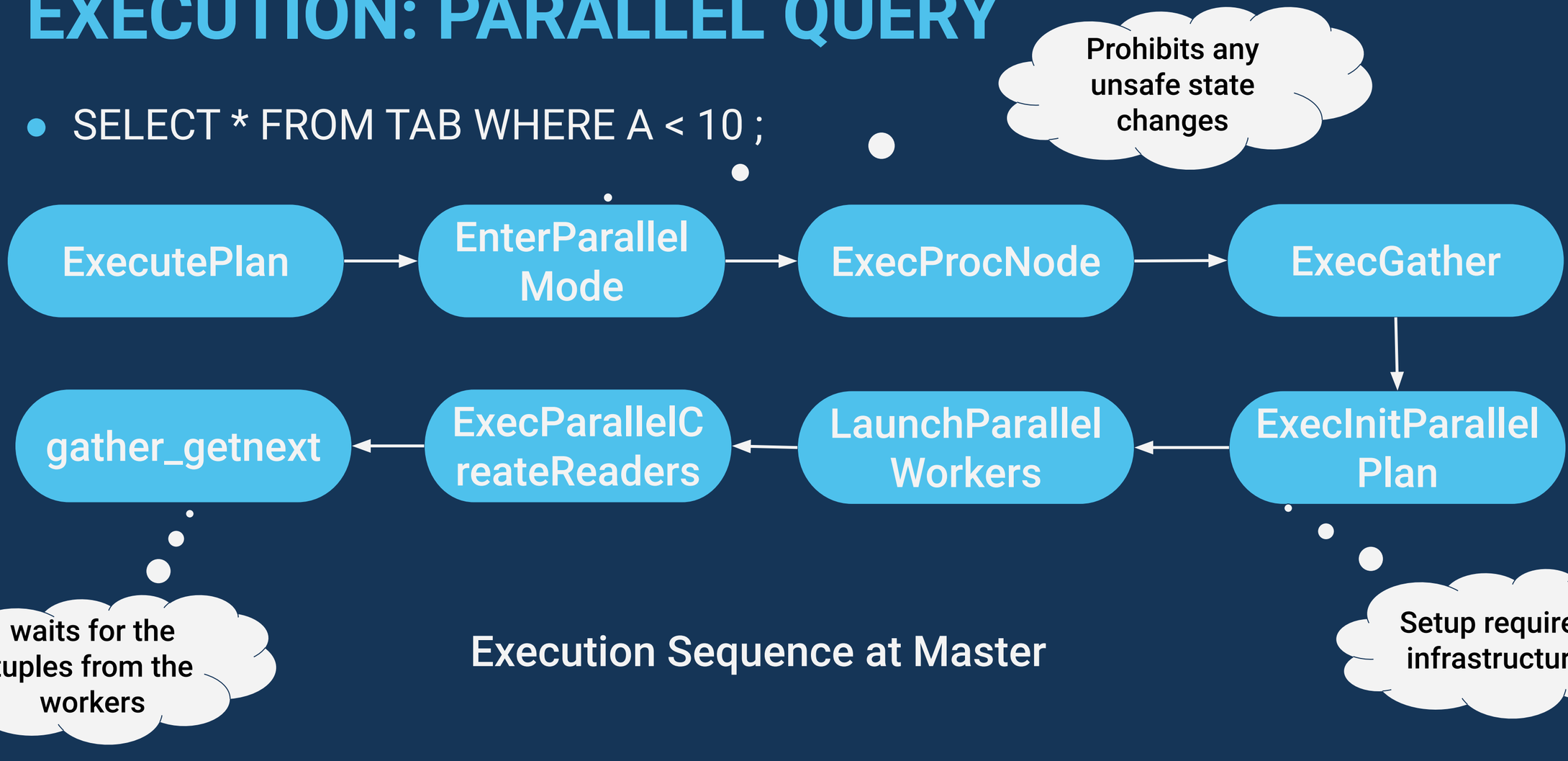
- SELECT * FROM TAB WHERE A < 10 ;

```
1| Gather  (cost=1000.00..94737.85 rows=53 width=8) (actual time=0.673.. rows=19 loops=1)
2|  Workers Planned: 2
3|  Workers Launched: 2
4|  ->  Parallel Seq Scan on tab  (cost=0.00..93737.85 rows=22 width=8) (actual
   time=38.989..113.687 rows=6 loops=3)
5|        Filter: (a < 10)
6|        Rows Removed by Filter: 3167530
7| Planning Time: 0.216 ms
8| Execution Time: 123.700 ms
9| (8 rows)
```

JOURNEY OF THE QUERY

EXECUTION: PARALLEL QUERY

- SELECT * FROM TAB WHERE A < 10 ;

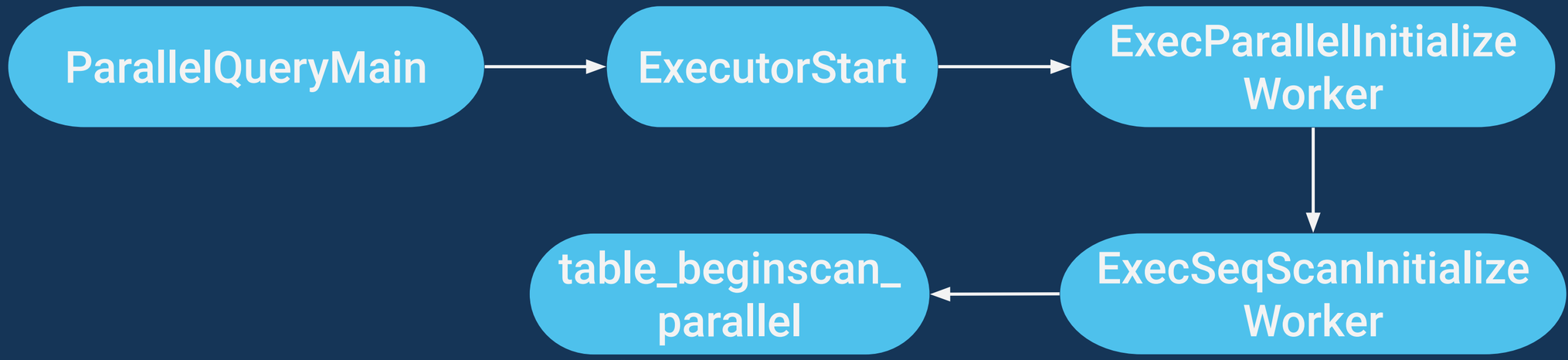


Execution Sequence at Master

JOURNEY OF THE QUERY

EXECUTION: PARALLEL QUERY

- SELECT * FROM TAB WHERE A < 10 ;



Initialize PlanState etc. based on shared_memory

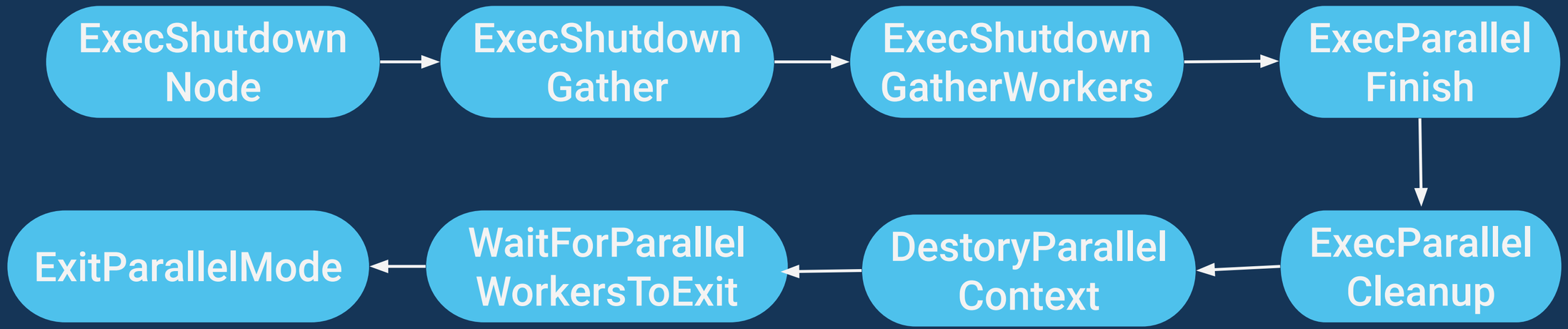
Parallel heap scan

Execution Sequence at Worker

JOURNEY OF THE QUERY

EXECUTION: PARALLEL QUERY

- SELECT * FROM TAB WHERE A < 10 ;



Execution Sequence at Master

JOURNEY OF THE QUERY

EXPRESSION EVALUATION

- In targetlist, where clauses, group by clauses, etc.
- Each separately executable expression tree is represented as a single ExprState node
- It contains the information to evaluate the expression in linear format
- ExprState
  - `struct ExprEvalStep *steps`
  - `ExprStateEvalFunc evalfunc`
  - `Expr    *expr`
- ExprEvalStep
  - intptr_t opcode
  - Datum   *resvalue
  - based on the instruction type, different inline structures are there

JOURNEY OF THE QUERY

EXPRESSION EVALUATION

- ExecInitExpr:
  - converts the Expr node tree to ExprState
  - precompute information if possible
  - each member of this array is of type ExprEvalSteps
  - it is non recursive
- ExecEndExpr
  - there is no such function
  - the memory is released with the reset/ delete of the memory context

JOURNEY OF THE QUERY

CLEANUP PHASE

- PortalDrop
  - ExecutorEnd
  - ExecEndNode
  - FreeExecutorState
    - Frees up all memory allocated for the query
  - FreeQueryDesc

- Drop respective buffer pins
- Close open relations

JOURNEY OF THE QUERY

EXECUTOR: REPO OVERVIEW

- Every exec node have their respective functions defined in respective files
  - scans - seq, index, bitmap (execScan, nodeBitmapHeapScan, …)
  - joins - nested loop, hash, merge
  - others - aggregate, sort, etc.

- There is a respective Init function to initialise the node to make necessary preparations
  - ExecInitSeqScan, ExecInitMergeJoin

JOURNEY OF THE QUERY

EXECUTOR: REPO OVERVIEW

- There are a few Exec functions for the respective node, to do the actual execution
  - ExecSeqScan, ExecSeqScanNext, ExecInsert
- There is a an end function to release the allocated storage
  - ExecEndSort, ExecEndAgg

JOURNEY OF THE QUERY

MEMORY MANAGEMENT

- All of the memory allocation in PostgreSQL is done via MemoryContext
- MemoryContexts are arranged as a forest
  - each context can have multiple children
  - each context can have maximum one parent
  - Reset/delete of a context causes its children also to reset/delete

JOURNEY OF THE QUERY

MEMORY MANAGEMENT

- The basic operations of a context are,
  - context creation
  - allocating memory
  - delete context
  - reset context
  - inquire about the total memory allocated in a context

- CurrentMemoryContext information available as a global variable

JOURNEY OF THE QUERY

MEMORY MANAGEMENT

- Some important MemoryContexts are
  - TopMemoryContext
  - PostmasterContext
  - CacheMemoryContext
  - TopTransactionContext
  - CurTransactionContext
  - ErrorContext

- A per-query memory context is created in CreateExecutorState()
- Most processing is done in per-tuple context to avoid intra-query memory leaks

JOURNEY OF THE QUERY

CONCLUSION

- CreateQueryDesc
  - ExecutorStart
    - CreateExecutorState – creates per-query context
    - AfterTriggerBeginQuery
    - ExecInitNode --- recursively scans plan tree
      - CreateExprContext – creates per-tuple context
      - ExecInitExpr

JOURNEY OF THE QUERY

CONCLUSION

- ExecutorRun
  - ExecProcNode --- recursively called in per-query context
    - ExecEvalExpr --- called in per-tuple context
    - ResetExprContext --- to free memory

- ExecutorFinish
  - ExecPostprocessPlan --- run any unfinished ModifyTable nodes
  - AfterTriggerEndQuery

JOURNEY OF THE QUERY

CONCLUSION

- ExecutorEnd
  - ExecEndNode --- recursively releases resources
  - FreeExecutorState – frees per-query context and child contexts

- FreeQueryDesc

KEEP EXECUTING!

THANK YOU !

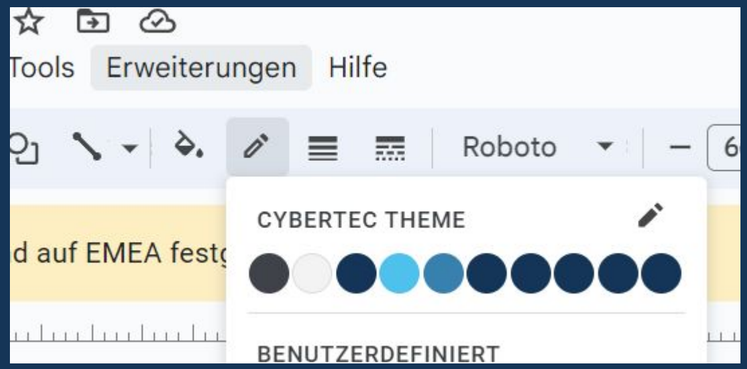
IMPORTANT DATA STRUCTURES

- Plan tree
- Presentations are communication tools that can be used as lectures, reports, and more.
- Presentations are communication tools that can be used as lectures, reports, and more.

GUIDE

SLIDES STYLE GUIDELINES

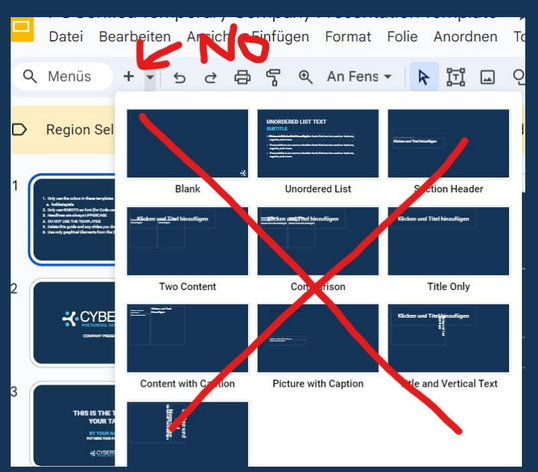
1. Only use the Colors from CYBERTEC THEME →
2. Only use Roboto normal as font
3. Only use Consolas bold for Code
4. HEADLINES ALWAYS IN UPPERCASE

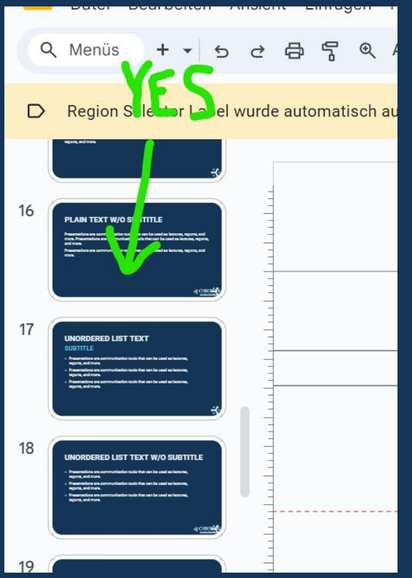


GUIDE

SLIDES USAGE GUIDELINES

1. DO NOT USE THE TEMPLATES: use the prepared slides on the left and copy them) →
2. For image slides use ONLY the template image slides (no full slide images)
3. Use only graphical Elements from Page 4
   (if you need other, get in touch with marketing)
4. **Delete this guide and any slides you don't need**





GUIDE

SLIDES INDEX

TEMPLATES

GUIDE

GRAPHICAL ELEMENTS & ICONS

This is a box with information

ÜBER CYBERTEC

Hoch spezialisiertes, schnell wachsendes IT Unternehmen

Internationales Team (10 Länder), weltweit 6 Standorte

Datenbank-, Data & Science Services

Inhabergeführt seit dem Jahr 2000


AUSTRIA (HQ)
CYBERTEC POSTGRESQL
INTERNATIONAL (HQ)
ESTONIA
CYBERTEC POSTGRESQL
NORDIC
SWITZERLAND
CYBERTEC POSTGRESQL SWITZERLAND
POLAND
CYBERTEC POSTGRESQL
POLAND
URUGUAY
CYBERTEC POSTGRESQL
SOUTH AMERICA
SOUTH AFRICA
CYBERTEC POSTGRESQL
SOUTH AFRICA

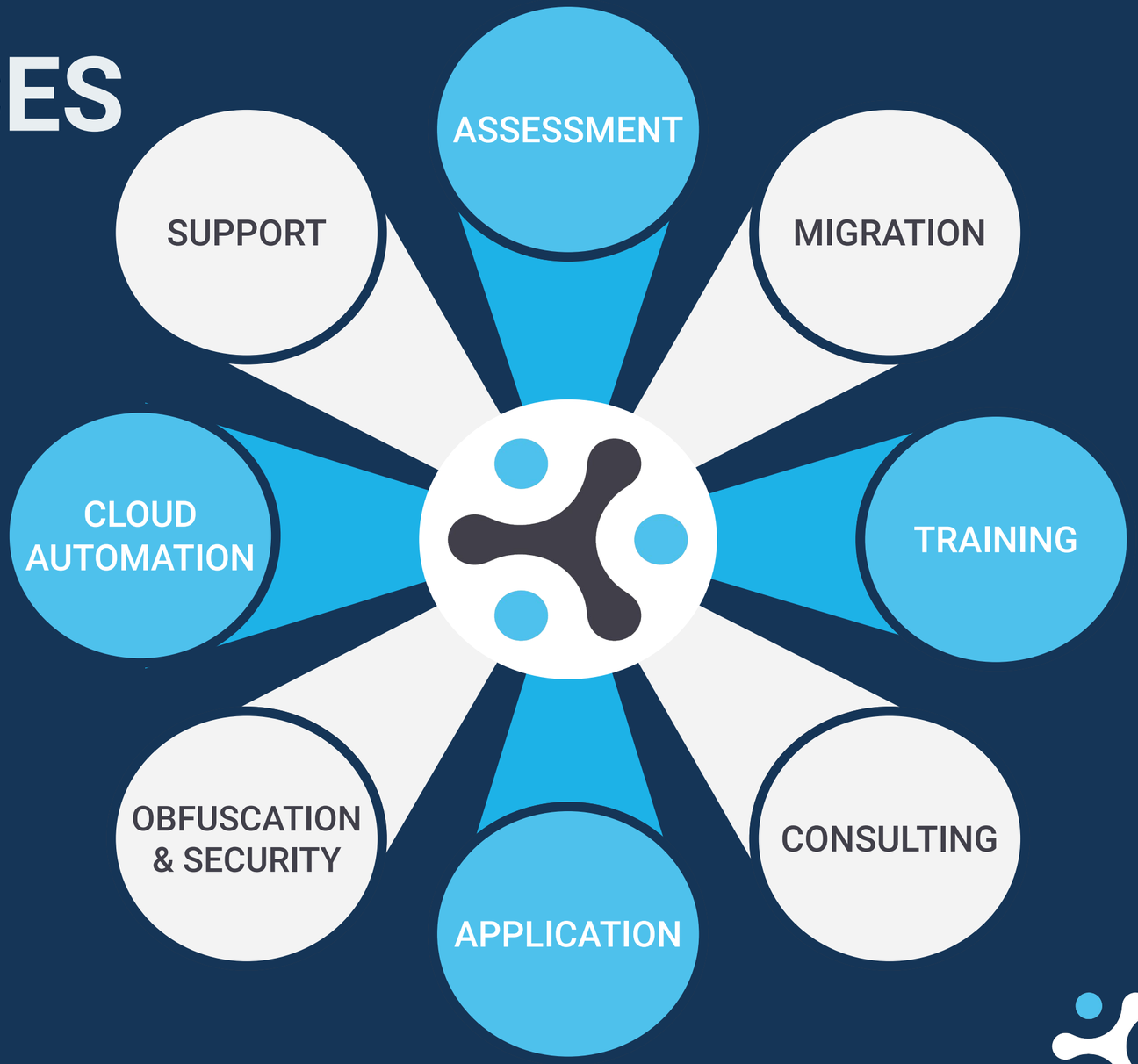
WARUM PostgreSQL?

www.cybertec-postgresql.com

WHY PostgreSQL?

DATABASE SERVICES

- 24/7 Support
- High Availability
- Consulting
- Performance Tuning
- Clustering
- Migration
- Etc.



CODE SLIDE

```
query = """SELECT DISTINCT *
             FROM (
                  SELECT sources.id, sources.name FROM sources
                  WHERE sources.suite='{suite}' AND sources.architecture='{arch}'
                  AND sources.id NOT IN
                     (SELECT schedule.package_id FROM schedule WHERE
build_type='ci_build')
                  AND sources.id NOT IN
                     (SELECT results.package_id FROM results)
                  ORDER BY random()
              ) AS tmp
              LIMIT {limit}""".format(suite=suite, arch=arch, limit=limit)
```

PLAIN TEXT

SUBTITLE

Presentations are communication tools that can be used as lectures, reports, and more. Presentations are communication tools that can be used as lectures, reports, and more.

Presentations are communication tools that can be used as lectures, reports, and more.

UNORDERED LIST W/O SUBTITLE

- Presentations are communication tools that can be used as lectures, reports, and more.
- Presentations are communication tools that can be used as lectures, reports, and more.
- Presentations are communication tools that can be used as lectures, reports, and more.

PLAIN TEXT W/O SUBTITLE

Presentations are communication tools that can be used as lectures, reports, and more. Presentations are communication tools that can be used as lectures, reports, and more.

Presentations are communication tools that can be used as lectures, reports, and more.

UNORDERED LIST TEXT

SUBTITLE

- Presentations are communication tools that can be used as lectures, reports, and more.
- Presentations are communication tools that can be used as lectures, reports, and more.
- Presentations are communication tools that can be used as lectures, reports, and more.

ORDERED LIST TEXT

SUBTITLE

1. Presentations are communication tools that can be used as lectures, reports, and more.
2. Presentations are communication tools that can be used as lectures, reports, and more.
3. Presentations are communication tools that can be used as lectures, reports, and more.

ORDERED LIST W/O SUBTITLE

1. Presentations are communication tools that can be used as lectures, reports, and more.
2. Presentations are communication tools that can be used as lectures, reports, and more.
3. Presentations are communication tools that can be used as lectures, reports, and more.

IMAGE PLACEHOLDER

-

1. DOUBLE-CLICK THIS IMAGE
2. DRAG AND DROP THE NEW IMAGE HERE
3. POSITION IT CORRECTLY

OUR WORLD

TECHNOLOGY AROUND THE GLOBE

Presentations are communication tools that can be used as demonstrations, lectures, speeches, reports, and more. Most of the time, they’re presented before an audience. It serves a variety of purposes, making them powerful tools for convincing and teaching.

IMAGE PLACEHOLDER

-

1. DOUBLE-CLICK THIS IMAGE
2. DRAG AND DROP THE NEW IMAGE HERE
3. POSITION IT CORRECTLY

OUR WORLD

TECHNOLOGY AROUND THE GLOBE

Presentations are communication tools that can be used as demonstrations, lectures, speeches, reports, and more.

OUR WORLD

TECHNOLOGY AROUND THE GLOBE

Presentations are communication tools that can be used as demonstrations, lectures, speeches, reports, and more. Most of the time, they’re presented before an audience. It serves a variety of purposes, making them powerful tools for convincing and teaching.

OUR WORLD

TECHNOLOGY AROUND THE GLOBE

Presentations are communication tools that can be used as demonstrations, lectures, speeches, reports, and more.

IMAGE PLACEHOLDER

-

1. DOUBLE-CLICK THIS IMAGE
2. DRAG AND DROP THE NEW IMAGE HERE
3. POSITION IT CORRECTLY

Pre-Digital

Presentations are communication tools that can be used as lectures.

IMAGE PLACEHOLDER

-

1. DOUBLE-CLICK THIS IMAGE
2. DRAG AND DROP THE NEW IMAGE HERE
3. POSITION IT CORRECTLY

Post-Digital

Presentations are communication tools that can be used as lectures.

WHERE DO WE GO NEXT?

HOW DO WE GET THERE?

“THIS IS A BRAND NEW QUOTE, USE IT OR LOSE IT :)”

ALBERT EINSTEIN

HANS-JÜRGEN SCHÖNIG

CEO & FOUNDER

EMAIL

hs@cybertec-postgresql.com

PHONE

+43 2622 930 22 - 666

WEB

www.cybertec-postgresql.com

NAME & SURNAME

YOUR POSITION

EMAIL

EMAILXXX@cybertec-postgresql.com

PHONE

Your Phone Number XXX

WEB

www.cybertec-postgresql.com

IMAGE PLACEHOLDER

-

1. DOUBLE-CLICK THIS IMAGE
2. DRAG AND DROP THE NEW IMAGE HERE
3. POSITION IT CORRECTLY

CREATION OF TECHNOLOGY

PLANNING

Presentations are communication tools that can be used as lectures, reports, and more.

PLANNING

Presentations are communication tools that can be used as lectures, reports, and more.

PLANNING

Presentations are communication tools that can be used as lectures, reports, and more.

ADDITIONAL READING

TECH FUTURE TODAY
www.reallygreatsite.com

ADVANCES IN TECHNOLOGY
www.reallygreatsite.com

INNOVATIONS AND INVENTIONS
www.reallygreatsite.com

TIMELINE

BY MILLENIUM

2ND MILLENNIUM (ABC)

Presentations are communication tools that can be used as lectures.

1ST MILLENNIUM (BC)

Presentations are communication tools that can be used as lectures.

1ST MILLENNIUM (AD)

Presentations are communication tools that can be used as lectures.

2ND MILLENNIUM (AD)

Presentations are communication tools that can be used as lectures.

3RD MILLENNIUM (AD)

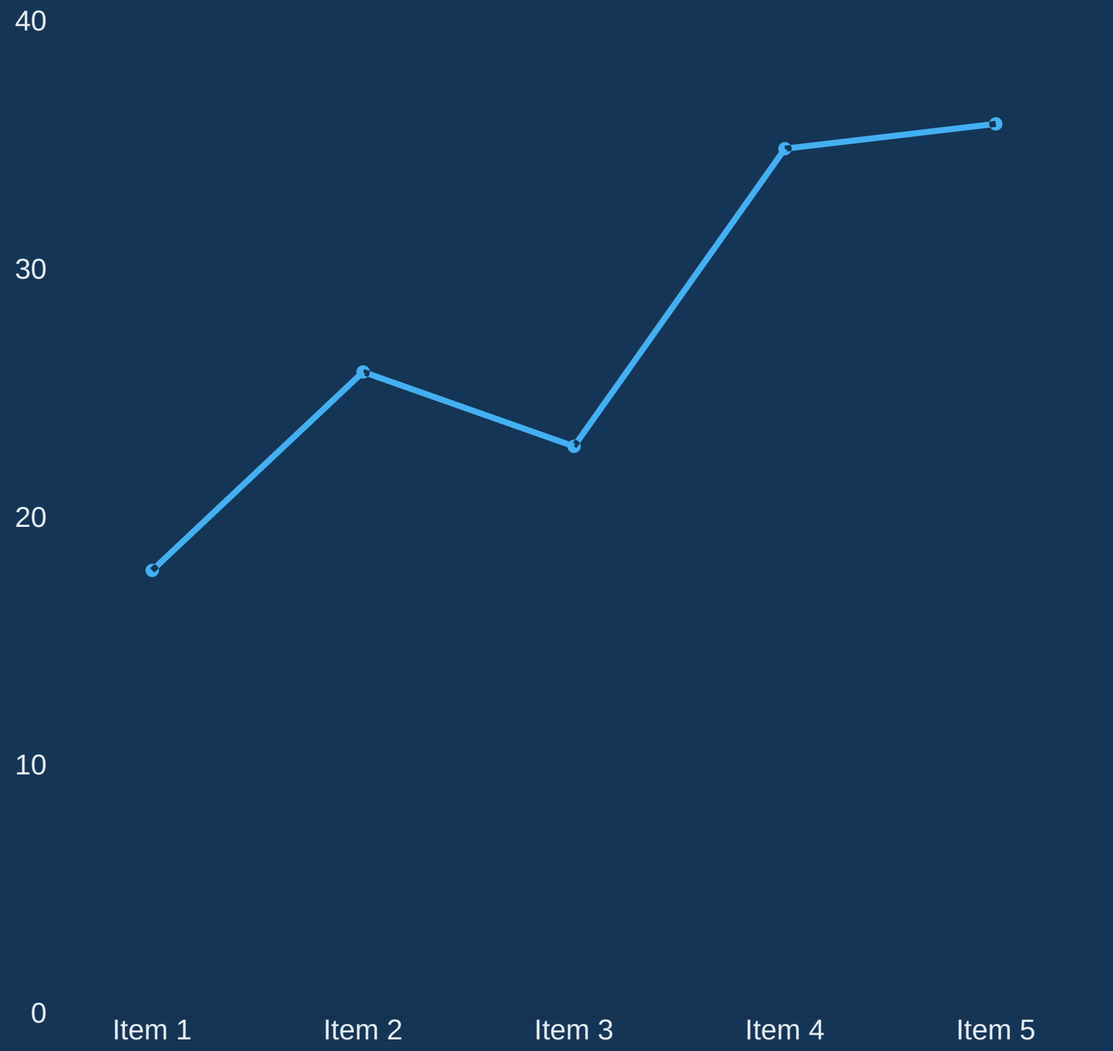
Presentations are communication tools that can be used as lectures.

Relationship with Technology

94% OF STUDENTS

USE THEIR SMARTPHONES EVERY HOUR

Presentations are communication tools that can be used as lectures.



NUMBER OF DEVICES

BY AGE GROUP

Presentations are communication tools that can be used as demonstrations, lectures, srpeeches, reports, and more.

USE OF TECHNOLOGY

PERSONAL USE

Presentations are tools that can be used as lectures.

COMMUNITY USE

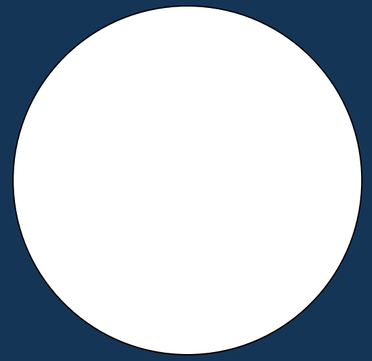
Presentations are tools that can be used as lectures.

GLOBAL USE

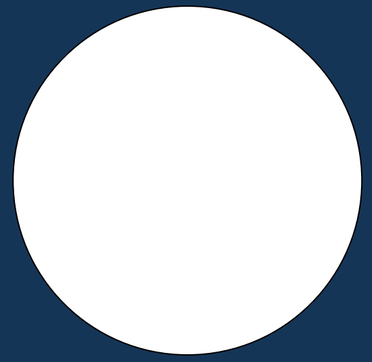
Presentations are tools that can be used as lectures.

GROUP 3 MEMBERS

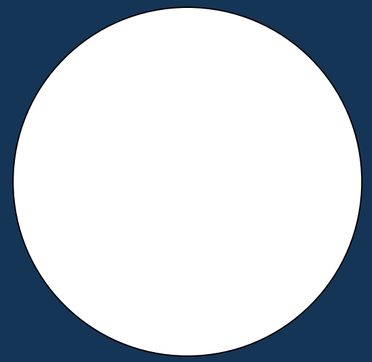
MEET OUR TEAM



MITCHELL TRINIDAD
Group Speaker



NANETTE PRESTON
Group Leader



HANNAH REMINGTON
Lead Researcher