# Parser

## Úvod

Tento dokument opisuje udaje a process parseru.

## Údaje

### Tabuľka surového zdroja (RAWSRC)

Vstupom do parsera je zoznam otypovaných tokenov, čo znamená inými slovami tabuľka:

|  |  |  |
| --- | --- | --- |
| Index | Typ tokenu (text) | Token (text) |
| INDEX | TYPE | VALUE |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

INDEX je jedinečný a určuje poradie symbolov ako takých.

TYPE určuje typ tokenu. Pri čítaní textového vstupu je textový súbor rozďeľovaný do textových častí rôznej dĺžky. Každej takejto časti môže byť pridelený predbežne jej typ, ktorý určuje povahu znakov v tejto časti. Napríklad keď sú tam len číselné znaky, tak typ je číslo. Keď sú tam len čiarky, tak typ by mohol byť “oddelovač”

VALUE je textová hodnota danej textovej časti.

### Vstup - tabuľka zdroja (SRC)

Tabuľka zdroja má tú istú štruktúru aj význam ako RAWSRC. Rozdiel spočíva v tom, že z RAWSRC sú odfiltrované niektoré nežiadúce prvky ako sú napríklad medzery, nové riadky, komentáre a podobne. Do algoritmu parsera tak vstupujú len žiadúce elementy.

### Tabuľka symbolov (SYMBOLS)

Štruktúra tabuľky je nasledovná:

|  |  |
| --- | --- |
| ID (číslo) | SYMBOL (text) |
|  |  |

ID je jedinečné pre celú tabuľku.

SYMBOL je textová reprezentácia symbolu.

### Tabuľka definícií symbolov (SYMBDEF)

Je to konfiguračná tabuľka, kde sú definované rozvoje pre jednotlivé symboly. Táto tabuľka slúži jednak na identifikáciu tokenov zo vstupu ako aj identifikáciu iných symbolov z tejto tabuľky. Preto jeden symbol je definovaný postupnosťou elementov ktoré sú jedným z týchto 3 druhov:

* Element definovaný hodnotou
* Element definovaný typom
* Element ako odkaz na iný symbol

Štruktúra tabuľky je nasledovná:

|  |  |  |
| --- | --- | --- |
| SYMBOL\_DEF\_ID | SYMBOL\_ID | DEF[] |
| 1 | 18 |  |
| 2 | 18 |  |
| 3 | 20 |  |
|  |  |  |

Každý riadok v SYMBDEF má ešte pole, kde je definovaný zoznam elementov:

|  |  |  |
| --- | --- | --- |
| INDEX | TYPE | VALUE |
|  |  |  |

SYMBOL\_DEF\_ID je jedinečné pre celú tabuľku

SYMBOL\_ID je index do tabuľky symbolov

DEF[] je zoznam elementov reozvoja symbolu.

Prvky tohoto zoznamu majú svoje presné číslo v poradí, majú svoj typ a textovú hodnotu. Typ určuje, ako sa interpretuje a púoužije textová hodnota v stĺpci VALUE.

Sú tri typy elementov:

* Element definovaný hodnotou BY\_VALUE
* Element definovaný typom BY\_TYPE
* Element ako odkaz na iný symbol BY\_SYMBOL

DEF\_LENGTH je počet prvkov v zozname elementov DEF

### Zoznam definícií podľa prvého elementu - FIRSTDEF

Táto tabuľka je vytvorená na základe SYMBDEF tabuľky. Pri jej tvorbe sa zistí prvý element v konkrétnej definícii symbolu.

Má stromovú štruktúru pre rýchle prehľadávanie. Na prvej úrovni v strome sa rozlišuje typ elementu, takže tu máme 3 uzly: BY\_TYPE, BY\_VALUE, BY\_SYMBOL

Na druhej úrovni je uvedená textová hodnota elementu, ktorý je prvý v danej definícii symbolu. Táto textová hodnota je naviazaná na zoznam konkrétnych definícií symbolu identifikovanú svojim ID (alebo ak je tento strom v pamäti, tak obsahuje referenciu na danú definíciu).

FIRSTDEF

* BY\_TYPE
  + Ident
    - Symbdef 18
    - Symbdef 20
    - …
  + Punc
    - Symbdef 19
    - Symbdef 42
  + Word
    - Symbdef 17
* BY\_VALUE
  + +=
    - Symbdef 192
    - Symbdef 72
  + /=
    - Symbdef 34
    - Symbdef 25
    - …
* BY\_SYMBOL
  + ABC
    - Symbdef 36
    - …
  + CDE
    - Symbdef 76
    - Symbdef 38
    - …

### Tabuľka sledovania symbolov TRACKSYMB (skrátene TS)

Jednotlivé prvky sú vytvárané a vymazávané dynamicky v priebehu spracovávania vstupu.

**Vytvorenie prvku v TS**

Vytvára sa vtedy, keď sa vytváral nový záznam v T a ešte neexistoval jeho rodičovský TS (to znamená, že neexistoval TS kde jeho SYMBOL a POSITION boli zhodné s práve vytvoreným záznamom v T

**Zmazanie prvku z TS**

**Štruktúra tabuľky**

|  |  |  |
| --- | --- | --- |
| Atribút | Údajový typ | Chovanie |
| SYMBOL\_ID | Číslo, je to index do SYMBOLS | Inicializuje sa pri vytváraní a preberie sa hodnota z prvku T, na základe ktorého tento TS vytvárame. |
| POSITION | Číslo, Index do SRC | Inicializuje sa pri vytváraní a preberie sa hodnota z prvku T, na základe ktorého tento TS vytvárame. |
| END | Číslo, Index do SRC | Pri inicializácii má hodnotu -1.  Ak tento TS bol daný do stavu MATCHED, tak hodnotu END prevezmeme z matched elementu T, na základe ktorého sme dali aj tento TS do stavu MATCHED. Ak je viacero T matched, potom vyberáme ten T, ktorý má najväčšiu hodnotu END. |
| STATUS | Enum: PENDING, MATCHED, FAILED | Pri inicializácii sa nastavuje hodnota PENDING.  Ak sú všetky T pre daný SYMBOL\_ID a POSITION v stave FAILED, tak aj tento rodičovský TS sa nastaví do stavu FAILED.  Ak žiaden T pre daný symbol SYMBOL\_ID a POSITION nie je PENDING a aspoň jeden je MATCHED, tak aj tento rodičovský TS sa nastaví do stavu MATCHED. |

### Tabuľka sledovania konkrétnej definície symbolu TRACKS, skrátene T

Jednotlivé prvky sú vytvárané a vymazávané dynamicky v priebehu spracovávania vstupu.

**Vytvorenie prvku v T**

Vytvorí sa v jednom z týchto prípadov:

* Na vstupe zo SRC sa prečítal nový element a jeho typ a/alebo hodnota je prvým elementom v definícii SYMBDEF, vtedy sa pre daný SYMBOL\_ID a SYMBOL\_DEF\_ID vytvorí prvok v T.
* Keď nejaký TS bol daný do stavu matched, tak sa vytvoria také T, kde daný TS je na prvom mieste v definícii SYMBDEF okrem tých, ktoré patria pod seba samých.
* Keď nejaký T bol daný do stavu MATCHED, tak sa vytvoria také T, ktoré patria pod ten istý symbol a daný symbol je na prvom mieste ich SYMBDEF definíícii (čiže rekurzívne definície symbolov)

**Zmazanie prvku z T**

Ak rodičovský TS bol daný do stavu FAILED, tak všetky pridružené T môžu byť zmazané.

**Štruktúra tabuľky**

|  |  |  |
| --- | --- | --- |
| Atribút | Údajový typ | Chovanie |
| SYMBOL\_ID | Číslo | Je inicializovaný pri vytváraní. |
| SYMBOL\_DEF\_ID | Číslo, ID do tabuľky SYMBDEF | Je inicializovaný pri vytváraní. |
| POSITION | Ćíslo | Je inicializovaný pri vytváraní podľa druhu vytvárania, takže buď je to index do SRC, ak bol T vytvorený na zaáklade neho, alebo je to prevzatá hodnota POSITION z iného TS alebo T, na základe ktorého je tento T vytvorený |
| END | Ćíslo | Default hodnota je -1.  Na platný index do SRC sa nastavuje vtedy, keď posledný TP z TPS bol matched na základe toho elementu zo SRC. |
| STATUS | Enum: PENDING, MATCHED, FAILED | PENDING  Toto je defaultný stav.  MATCHED  Hneď ako sú všetky TP označené ako MATCHED, tak aj tento rodičovský T je označený ako MATCHED.  FAILED  Hneď ako je niektorí TP označený ako FAILED, tak aj tento celý T je označený ako FAILED. |
| TPS | Usporiadaný zoznam prvkov TP | Tento zoznam je vytvorený podľa SYMBDEF tabuľky pri vytváraní T. Zoznam samotný sa potom už nemení, menia sa už len atribúty jeho prvkov. |

### Tabuľka sledovania rozvoja symbolu TRACKPARTS, skrátene TP

Jednotlivé prvky sú vytvárané a vymazávané dynamicky v priebehu spracovávania vstupu.

**Vytvorenie prvku v TP**

Vytvára sa vtedy, keď sa vytvára rodičovský T a sú to len prekopírované položky zo SYMBDEF

**Zmazanie prvku z TP**

TP sa odstraňujú vtedy, ked je ich rodičovský T zmazaný

**Štruktúra tabuľky**

|  |  |  |
| --- | --- | --- |
| Atribút | Údajový typ | Chovanie |
| INDEX | Číslo (nemusí byť explicitne uvedené ako súčasť fyzickej štruktúry) | Obsahuje hodnoty vzostupne 0, 1, 2, 3, … podľa poradia pridávania týchto prvkov do poľa TPS |
| SYMBOL\_TYPE | Enum: BY\_TYPE, BY\_VALUE, BY\_SYMBOL | Hodnota prevzatá z tabuľky SYMBDEF.  BY\_TYPE: obsah SYMBOL\_VALUE bude porovnávaný so SRC atribútom TYPE  BY\_VALUE: obsah SYMBOL\_VALUE bude porovnávaný so SRC atribútom VALUE  BY\_SYMBOL: obsah SYMBOL\_VALUE bude porovnávaný s TS atribútom SYMBOL (v TS je len SYMBOL\_ID, ale cez tabuľku SYMBOLS dostaneme jeho textovú podobu)  Pre všetky 3 porovnávania je samozrejme rozhodujúca hodnota POSITION, ktorá musí byť zhodná buď s indexom do SRC, alebo POSITIOn v TS (podľa toho, ktorý typ porovnávania sme robili) |
| SYMBOL\_VALUE | text | Hodnota prevzatá z tabuľky SYMBDEF |
| POSITION | Číslo | Pri inicializácii prvý TP hodnotu prevzatú z rodičovského T, ostatné majú hodnotu -1.  Na novú hodnotu sa nastaví vtedy, keď prvok pred ním TP\_PREV bol práve daný do stavu MATCHED, vtedy vezme hodnotu TP\_PREV -> END, pripočíta hodnotu 1 a priradí ju do TP -> POSITION.  Ak POSITION nie je -1 a END je -1, tak tento TP je v stave PENDING a očakáva, že na pozícii POSITION v SRC sa vyskytne tento symbol definovaný hodnotou alebo typom (závisí od SYMBOL\_TYPE) uvedeným v SYMBOL\_VALUE, alebo že na pozícii POSITION sa v TS vyskytne symbol definovaný menom v SYMBOL\_VALUE. |
| END | Číslo | Pri inicializácii má hodnotu -1.  Ak je prvok daný do stavu MATCHED na základe prvku zo SRC, tak END nastaví na hodnotu indexu toho prvku zo SRC.  Ak je prvok daný do stavu MATCHED na základe prvku z TS, tak END prevezme z hodnoty END tohoto TS prvku. |
| STATUS | Enum | **PENDING**: Default hodnota  **MATCHED**: Ak tento prvok TP je zhodný s prečítaným prvkom SRC, alebo je zhodný s uzavretým TS na tejto pozícii POSITION.  **FAILED:** Ak je tento prvok určený typom alebo hodnotou a na vstupe SRC sa vyskytol zjavne iný prvok, potom je tento TP daný do stavu FAILED. Tu je nutné pripomenúť že ak tento TP očakáva prvok daný menom symbolu, potom očakáva len túto zhodu s tabuľkou TS a vstup zo SRC si nevšíma, pretože očakávaný symbol môže (ale nemusí) byť matched neskôr. |

### Zoznam pozícií LISTPOS

Je to zoznam symbolov, ktoré boli alebo sú aktívne, ktoré majú štartovaciu pozíciu na tomto indexe.

### Zoznam aktívnych definícií LISTDEF

Je to zoznam definícií symbolov, ktoré boli, alebo sú aktívne, ktoré majú štartovaciu pozíciu na tomto indexe.

### Zoznam aktívnych elementov LISTP

Je to zoznam pending elementov TP, ktoré očakávajú na MATCHED alebo FAILED, takže to znamená že sú v stave PENDING a že ich POSITION sa nerovná -1.